Best Design for Multidimensional Computerized Adaptive Testing With the Bifactor Model

Educational and Psychological

Measurement

2015, Vol. 75(6) 954–978

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415575147

epm.sagepub.com

Dong Gi Seo¹ and David J. Weiss²

Abstract

Most computerized adaptive tests (CATs) have been studied using the framework of unidimensional item response theory. However, many psychological variables are multidimensional and might benefit from using a multidimensional approach to CATs. This study investigated the accuracy, fidelity, and efficiency of a fully multidimensional CAT algorithm (MCAT) with a bifactor model using simulated data. Four item selection methods in MCAT were examined for three bifactor pattern designs using two multidimensional item response theory models. To compare MCAT item selection and estimation methods, a fixed test length was used. The D_s -optimality item selection improved θ estimates with respect to a general factor, and either D- or A-optimality improved estimates of the group factors in three bifactor pattern designs under two multidimensional item response theory models. The MCAT model without a guessing parameter functioned better than the MCAT model with a guessing parameter. The MAP (maximum a posteriori) estimation method provided more accurate θ estimates than the EAP (expected a posteriori) method under most conditions, and MAP showed lower observed standard errors than EAP under most conditions, except for a general factor condition using D_s -optimality item selection.

Keywords

computerized adaptive testing, multidimensional item response theory, bifactor model, full information item factor analysis

¹National Registry of Emergency Medical Technicians, Columbus, OH, USA

²University of Minnesota, Minneapolis, MN, USA

Corresponding Author:

Dong Gi Seo, National Registry of Emergency Medical Technicians, 6610 Busch Boulevard, Columbus, OH 43229, USA.

Email: dseo@nremt.org

Introduction

In the past several decades, research has repeatedly demonstrated that a computerized adaptive test (CAT) can be at least an average of 50% shorter than a paper-and-pencil test with equal or better measurement precision (Chang & van der Linden, 2003; Gibbons et al., 2008; Kingsbury & Weiss, 1980, 1983; van der Linden, 1998; Weiss, 1982, 1985). However, CAT primarily has been studied using the framework of unidimensional item response theory (UIRT), which is widely used in educational and psychological research to model how examinees respond to test items. Although most item response theory (IRT) models that are currently used assume that test items measure a single dominant latent trait, it is not always practical to assume that a test measures only a single latent trait (Reise, Morizot, & Hays, 2007). Most personality inventories in psychology are designed to measure multidimensional latent traits rather than a single latent trait (e.g., the NEO Personality Inventory; Costa & McCrae, 1992). Therefore, it is appropriate to introduce a multidimensional latent space, in which multidimensional IRT (MIRT) modeling would be adopted beyond the framework of unidimensionality in search of a more generalizable model to fit real data.

Several studies have examined the effects on item parameter estimation if UIRT is applied to multidimensional data (Ackerman, 1989; Ansley & Forsyth, 1985; Reckase, 1974). A general finding from these studies is that if there is a predominant general factor in the data, the presence of multidimensionality has little effect on the estimation of item and trait parameters. However, if the data have strong secondary factors beyond the primary factor, the application of a UIRT model results in a serious distortion of the measurement characteristics of the instrument. Indeed, the validity of UIRT applications (linking, model-fit, parameter estimation, scoring, and CAT) would also be questioned in a situation in which it is reasonable to hypothesize a multidimensional latent space. Consequently, CAT might not guarantee an optimal test for individual examinees unless IRT parameter estimates are accurately prespecified given an appropriate model for the data.

Through both adaptive and conventional testing with dichotomous three-parameter logistic (3PL) IRT, Folk and Green (1989) demonstrated that a unidimensional model applied to two-dimensional data affects item parameter estimates. Their study demonstrated that if nondominant factors do not affect scale scores, the trait (θ) can be estimated with an assumption of a unidimensional latent trait underlying the data in conventional testing. However, if the two-dimensional latent traits are relatively uncorrelated and dominant in the data, using one or the other trait creates a large difference in θ estimates. Folk and Green (1989) concluded that the difference between θ estimates in CAT was greater than in conventional testing because UIRT item discrimination parameter estimates are used for both the item selection and θ estimation procedures in CAT.

Since Bock and Aitkin (1981) extended the IRT model to a multidimensional case, many researchers have studied CAT using a bank of items calibrated under MIRT models. Initially, Bloxom and Vale (1987) developed multidimensional

adaptive estimation procedures. They extended the multivariate analysis of Owen's (1975) sequential Bayesian adaptive updating algorithm. Then, Tam (1992) evaluated a multidimensional adaptive estimation procedure through precision, test information, and computational time. However, these studies (Bloxom & Vale, 1987; Tam, 1992) considered the implementation of CAT with respect to only θ estimation methods. They did not address the procedure for multidimensional adaptive item selection that considers prior knowledge of a multivariate distribution of θ .

Although Bloxom and Vale (1987) and Tam (1992) initially developed multidimensional CAT (MCAT), their MCAT failed to demonstrate advantages over unidimensional adaptive testing (Segall, 1996). Therefore, Segall (1996) developed multidimensional Bayesian item selection and θ estimation procedures and demonstrated that his MCAT was more efficient than a unidimensional CAT (UCAT) in terms of test length and precision. In addition to gaining efficiency, his MCAT could be used as an instrument to measure various content traits for examinees from nine different subtests of the Armed Services Vocational Aptitude Battery (Moreno & Segall, 1992). Luecht (1996) also demonstrated the efficiency of MCAT. He observed that an MCAT with content constraints could achieve approximately the same precision with 25% to 40% fewer items than were required in UCAT with regard to the measurement of latent traits.

Furthermore, Li and Schafer (2005) showed that UCAT and MCAT, with constraints on item exposure rates, were capable of producing accurate estimates of reading and math abilities. Specifically, compared with UCAT, MCAT slightly increased the accuracy of θ estimates for examinees at the low and high end of the θ scale in both reading and math tests. Therefore, MCAT appears to be an efficient method for ensuring adequate coverage of content in adaptive testing and provides a separate multidimensional vector of estimated θ s for each examinee.

Segall (1996) used a confirmatory simple structure to implement MCAT, in which items within one scale were assumed to measure the same latent trait, and each item was loaded on only one latent trait. For that reason, this confirmatory simple structure MIRT model is called a "multi-unidimensional" IRT (multi-UIRT) model or a "between-item" MIRT model (Wang & Chen, 2004). However, these models confine each item to measuring a single latent trait similar to the multiple scale procedure, which is not realistic for many multidimensional constructs. Usually, the constraint of dimensional independence among factors is not appropriate for correlated data because latent traits are generally correlated with each other. Although MIRT models (e.g., Ackerman, 1989; Bock & Aitkin, 1981; Browne, 2001) allow latent traits to correlate with each other, they do so from an exploratory factor analytic perspective so that their latent traits are not readily interpretable.

To eliminate this constraint, a multidimensional item response model is needed that (a) measures more than one latent trait, (b) yields readily interpretable latent traits, and (c) directly estimates item and person parameters jointly. As a response to the need, the bifactor model was applied in CAT (Weiss & Gibbons, 2007), and the second-order factor model also was used in CAT (Huang, Chen, & Wang, 2012).

Figure 1. Four types of IRT models based on a confirmatory factor analytic perspective.

Figure 1a describes the UIRT model, and Figure 1b represents the multi-UIRT model (Segall, 1996; Wang & Chen, 2004). However, these models have not been applied to empirical data analysis to investigate latent trait structures, such as intelligence (e.g., Horn, 1986). Therefore, many researchers have described the simple models by adding factors between the test-specific factors and the general factor. These models can be formulated within the framework of second-order factor analysis, which is similar to the multi-UIRT model. Figure 1c illustrates a path-analytic representation of a simple second-order factor model, involving six observed variables (X1-X6), two first-order factors (F1 and F2), and a general factor based on the correlation between these two factors. The effect of the general factor on the observed variables is mediated by a particular first-order factor, and the effect size is proportional to the loading of the first-order factor on the general factor. This second-order factor model is different from the preceding models that are characterized by group factors or a single broad construct. A major advantage of the second-order factor approach is to simultaneously identify first-order factors and second-order factors. Huang et al. (2012) have implemented the second-order IRT model in MCAT. However, second-

order factors are conceptually abstract constructs and have different interpretations from first-order factors because the higher-order factors are not directly related to observed variables (Chen, West, & Sousa, 2006). Additionally, the use of MCAT with the second-order IRT model was limited to the number of first-order latent traits because of the inefficiency and complexity of estimating parameters of the second-order IRT model (Huang et al., 2012). Figure 1d presents the bifactor model of the six observed variables, which is simply an extension of Spearman's two-factor model. In a theoretical framework, a general factor in the bifactor model contributes to all variables with group factors. Because a general factor and group factors are all first-order factors, it is not more complicated to apply the bifactor model in MCAT than UCAT. Additionally, the general factor can be interpreted as an essentially unidimensional trait in an IRT model if the general factor loadings are more dominant than the group factor loadings (Reise et al., 2007).

Bifactor Models

Holzinger and Swineford (1937) originally applied the term *bifactor* to a test measuring psychological traits. They defined the bifactor pattern as a theoretical framework in which all variables are explained by a general factor and group factors, both as first-order factors. This bifactor pattern assumes that uncorrelated group factors are independent of the general factor. The bifactor model allows only one of the $k = 2, \dots, p$ values of λ_{ik} (group factor loadings) to be nonzero, in addition to λ_{i1} (the general factor loading). For example, the theoretical bifactor pattern with one general factor and two group factors for six items can be described as

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & 0 \\ \lambda_{41} & 0 & \lambda_{43} \\ \lambda_{51} & 0 & \lambda_{53} \\ \lambda_{61} & 0 & \lambda_{63} \end{bmatrix} \quad (1)$$

The first column is the general factor, and the other columns are the group factors in the factor pattern matrix. Cai, Yang, and Hansen (2011) proposed a bifactor-like structure if an item always loads on the general factor and is permitted to load on at most one specific factor. For example, if Items 3 and 4 did not load on a specific group factor, λ_{32} and λ_{43} would be zero in Equation 1.

The Schmid-Leiman solution allows the bifactor model to build an unrestricted model (exploratory bifactor model) from a polychoric correlation matrix, which can be implemented in the R function "schmid" ("psych" package; Revelle, 2015), SAS or SPSS macros (Wolff & Preising, 2005). Target pattern rotation is another way to estimate a less restricted model for the bifactor model (Browne, 2001). A confirmatory bifactor model can be directly implemented by the R program ("sem" package), Mplus (Muthén & Muthén, 2004), and LISREL (Jöreskog & Sörbom, 1995). Jennrich

and Bentler (2012) introduced an exploratory bifactor analysis using a bifactor rotation criterion. However, those methods are based on a linear factor model using a factor correlation matrix. In these applications, the models are employed to obtain the dimensions only instead of examinees' scores.

Gibbons and Hedeker (1992) specified the full-information item bifactor analysis (FIIBFA) model as combining a bifactor model with the multi-UIRT model representing simple structure, meaning that each item is related to a general trait and one group trait only. In the two-dimensional computation in the bifactor model, the primary dimension should be considered first, and then the second dimension can be considered to estimate the probability of a correct response. Consequently, the conditional probability of the item response $u_{ij}=1$ in the FIIBFA model can be described as

$$P(u_{ij}=1|\theta_{j1}, \theta_{jk}, \lambda_{i1}, \lambda_{ik}\tau_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \int_{\tau_i}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{X_i - \lambda_{i1}\theta_{j1} - \lambda_{ik}\theta_{jk}}{\sigma_i}\right)^2\right) dX_i, \quad (2)$$

where the latent variable X_i is assumed to follow $N(0, 1)$, τ_i is the threshold of item i , the data are assumed to be sampled from a population of people whose θ s follow a particular multivariate distribution, and $\sigma_i = \sqrt{1 - \lambda_{i1}^2 - \lambda_{ik}^2}$. The unconditional probability in a FIIBFA model can be justified by evaluating the probability of dimensions $2, \dots, k$, after integrating with respect to the distribution of θ_1 . The bifactor restriction reduces the p -dimensional integral to a two-dimensional integral, one for θ_1 and the other for $\theta_2, \dots, \theta_k$, which alleviates the computational burden in estimating item and person parameters compared with usual MIRT models.

Chen et al. (2006) differentiated the bifactor model from a second-order factor model. A bifactor model is potentially applicable if there are multiple domain group factors, each of which is hypothesized to account for the unique influence of the specific domain over and above the general factor. However, a second-order factor model is potentially applicable if the first-order factors are substantially correlated with each other, and there is a second-order factor that is based on the correlations among the first-order factors. The second-order factor model can be employed to test whether a second-order factor explains the first-order factors.

Reise et al. (2007) demonstrated that 16 items in the consumer assessment of health care providers and systems (CAHPS 2.0) were fitted well both the bifactor model and a second-order factor model. However, they showed that the major part of common variance was explained by a general factor in the bifactor model, which is a main conceptual difference with the second-order factor in a second-order factor model. A second-order factor contains a qualitatively different dimension from first-order factors because a second-order factor explains the common variance among first-order factors, not observed variables. However, a general factor in the bifactor model is on the same conceptual level with group factors. Consequently, Reise et al. (2007) said that although both the bifactor model and second-order factor model can provide the same fit to data, the second-order factor model does not directly address

if the data is unidimensional or multidimensional, whereas the bifactor model can do so directly.

A number of researchers have demonstrated that the bifactor model provides an excellent framework to measure multidimensional traits containing a primary construct (Gustafsson & Aberg-Bengtsson, 2010; Reise, Moore, & Haviland, 2010). However, the bifactor model is still poorly understood and seldom used by applied researchers. Initially, Weiss and Gibbons (2007) implemented a CAT algorithm with the bifactor model using the completed 616 items of the 626 items in "The Mood-Anxiety Spectrum Scales" (Cassano et al., 1997) and evaluated the efficiency and precision of the performance of CAT with the bifactor model. However, Weiss and Gibbons' (2007) algorithm was still based on unidimensional testing. Therefore, there is a need for an algorithm of item selection and scoring that is truly multidimensional for CAT with a bifactor model. The objective of this study was to evaluate appropriate multidimensional item selection and θ estimation methods to implement an MCAT bifactor algorithm to estimate a general θ and group θ s. The efficiency and precision of the MCAT algorithms were investigated. To address this goal, a Monte Carlo study was conducted to evaluate MCAT with the bifactor model. Four multidimensional item selection algorithms and two θ estimation methods were compared under three bifactor pattern designs.

Method

Four factors that reflect realistic testing situations and could affect the precision of CAT were considered: (1) two MIRT models, (2) four item selection methods, (3) three bifactor pattern designs, and (4) two θ estimation methods. The comparison was based on three dependent variables, including the correlation between true θ and estimated θ ($\hat{\theta}$), root mean square error (RMSE), and observed standard error (OSE).

Response Generation

To approximate the condition of equal measurement precision throughout the θ range, item banks contained 400 dichotomous items for the bifactor model and 600 items for a "bifactor-like" model, which were approximately the numbers of items in the Mood-Anxiety Spectrum Scales for which Weiss and Gibbons (2007) implemented a CAT algorithm with the bifactor model. The item responses were generated according to the bifactor model using an R program (R Development Core Team, 2012). In the Monte Carlo simulation study, IRT parameters that could be transformed into factor analytic parameters were specified. The equation for the probability of a correct response for a 3PL bifactor IRT model is

$$P(u_{ij} = 1) = c_i + \frac{1 - c_i}{1 + \exp[-1.702 \times \mathbf{a}_i'(\boldsymbol{\theta}_j - \mathbf{b}_i \mathbf{1})]}, \quad (3)$$

where $\mathbf{a}$ indicates the vector of discrimination parameters of item i , θ_j indicates the vector of the general and group factor latent traits of examinee j , c_i is the guessing parameter for item i , and b_i is the difficulty parameter of item i . Similar to MIRT models, c_i was set to 0.20 for a multidimensional compensatory three-parameter logistic MIRT (3PL MIRT) model because the probability of a successful random guess can be computed as 1 divided by the number of options (this study assumed five options) and set at zero for a 2PL MIRT in this study.

The item responses for this study were generated given the true θ s and item parameters using Equation 3. The first step in the data generation process was to generate 400 or 600 random numbers from $U[0, 1]$ for each examinee. The probability of a correct response given the 2PL or 3PL bifactor IRT model was obtained for each item, conditional on θ . These model-based probabilities were compared with the random numbers to obtain the item responses for each item. If the model-based probability was greater than the random number, the response to that item was recorded as correct (1). Likewise, if the model-based probability was less than the random number, the item response was recorded as incorrect (0). This process was repeated for each item to obtain the full item response matrix for the 400 or 600 items for each simulated examinee. To reduce the variance of the dependent variables, a total of 1,000 simulees were generated within each of three sets of bifactor pattern designs.

Bifactor Pattern Designs

Each item was assigned a vector of discrimination parameters $\mathbf{a}_i$ corresponding to each factor (general factor and group factors). Each item was allowed to load on the general factor and a single group factor. Therefore, the vectors of discrimination parameters for items contained in the first group factor took the form $\mathbf{a}_i = \{a_{iG}, a_{ig1}, 0\}$, where $a_{iG} > 0$ and $a_{ig1} > 0$. Similarly, the vectors of discrimination parameters for items that loaded on the second group factor took the form $\mathbf{a}_i = \{a_{iG}, 0, a_{ig2}\}$. This study also included the bifactor-like pattern (Cai et al., 2011) that had items with one general factor and one group factor, and items that loaded on the only general factor took the form $\mathbf{a}_i = \{a_{iG}, 0, 0\}$.

A bifactor model is especially appropriate if researchers have instruments with a dominant general factor (Reise et al., 2007). Reise et al. (2007) stated that if items tend to have small loadings on the general factor and large loadings on the group factor, the multi-UIRT model should be used, and if the general factor loadings are larger than group factor loadings, the bifactor model should be used to measure traits. Table 1 shows three bifactor pattern designs to represent typical bifactor patterns and bifactor-like pattern. The purpose of the three pattern designs was to examine the effect of bifactor pattern on the estimates of general and group factors. All three pattern designs were reasonable when translated into factor loadings without any Heywood cases (Heywood, 1931), and all values were positive values. To generate standardized general and group factor discrimination parameters, values of the multidimensional discrimination index (MDISC; Reckase, 1985) were drawn from a

Table 1. Three Discrimination Parameter Conditions for the Bifactor Models.

Items	Bifactor low			Bifactor high			Bifactor-like		
	G	g_1	g_2	G	g_1	g_2	G	g_1	g_2
1-200	$\text{Cos}(\pi/12) \times \text{MDISC}$	$\text{Cos}(\pi/2.5) \times \text{MDISC}$	0	$\text{Cos}(\pi/6) \times \text{MDISC}$	$\text{Cos}(\pi/3) \times \text{MDISC}$	0	$\text{Cos}(\pi/6) \times \text{MDISC}$	$\text{Cos}(\pi/3) \times \text{MDISC}$	0
201-400	$\text{Cos}(\pi/12) \times \text{MDISC}$	0	$\text{Cos}(\pi/2.5) \times \text{MDISC}$	$\text{Cos}(\pi/6) \times \text{MDISC}$	0	$\text{Cos}(\pi/3) \times \text{MDISC}$	$\text{Cos}(\pi/6) \times \text{MDISC}$	0	$\text{Cos}(\pi/3) \times \text{MDISC}$
401-600	—	—	—	—	—	—	$\text{Cos}(\pi/6) \times \text{MDISC}$	0	0

Note. G = general factor; g_1 = first group factor; g_2 = second group factor; MDISC = multidimensional discrimination index.

log-normal distribution with a mean of zero and standard deviation of 0.20. In the traditional bifactor pattern with the low group factor discrimination parameters, the discrimination values were then calculated from MDISC such that the first 200 items had an angle of 15 degrees with the general factor axis and 75 degrees in the first group factor axis, whereas Items 201 to 400 had angles of 15 degrees with the general factor axis and 75 degrees in the second group factors. With the specified angles between the general and group factors, the general factor had some items that weighted more heavily in its direction in order to reduce the indeterminacy of which factor would likely be the general factor (DeMars, 2007). In the traditional bifactor pattern with the high group factor discrimination parameters, the discrimination values were then calculated from MDISC such that the first 200 items had an angle of 30 degrees with the general factor axis, and 60 degrees with the first group factor axis, whereas Items 201 to 400 had angles of 30 degrees with the general factor axis and 60 degrees with the second group factors. In the bifactor-like pattern, the discrimination values were then calculated from MDISC such that the first 200 items had an angle of 30 degrees with the general factor axis and 60 degrees with the first group factor axis, whereas Items 201 to 400 had angles of 30 degrees with the general factor axis and 60 degrees with the second group factor and Items 401 to 600 loaded on only the general factor with MDISC.

For an item bank providing equal measurement precision across θ s, item difficulty parameters, b_i , should be evenly and equally distributed throughout the θ continuum of interest (Weiss, 1982). Therefore, for each item bank, the b_i parameters were randomly generated from a uniform distribution from -4 to 4 so that they were equally distributed across θ s.

True θ s

Although the bifactor model can be used as an exploratory type of analysis using a bifactor rotation criterion (Jennrich & Bentler, 2012), this study applied a confirmatory bifactor analysis within an IRT framework. Because the bifactor model in this study was constructed so that uncorrelated group factors were independent of the general factor (Holzinger & Swineford, 1937), it was not necessary to consider the inter-correlations of traits. Therefore, each examinee had latent traits (θ) that were orthogonal to each other. Their θ s were randomly drawn from a multivariate normal distribution with $MVN(\mathbf{0}, \mathbf{I})$ without any correlations among θ . These true θ s were generated by the “mvtnorm” package (Genz & Bretz, 2009) in R for the bifactor model with one general factor and two group factors.

θ Estimation

The multidimensional maximum a posteriori (MAP; Bock & Aitkin, 1981) method was used to estimate θ for each examinee in the MCAT with the bifactor model. A standard multivariate normal distribution was used as the prior for MAP. MAP is a Bayesian method that incorporates information about the prior θ distribution to better

estimate the posterior distribution of θ (Bock & Mislevy, 1982). The MAP estimates of θ can be approximated by setting the partial derivative of the log of the posterior distribution to zero (Baker, 1992; Bock & Mislevy, 1982). The Newton-Raphson procedure was used to estimate θ for MAP. The Newton-Raphson iterations were repeated until the incremental change in $\hat{\theta}_s$ became less than the criterion of .001.

In the bifactor model, Gibbons et al. (2007) simplified the expected a posteriori (EAP) method to estimate the primary latent variable θ_1 and subdomain trait scores θ_k , given the observed response vector $\mathbf{u}_j$ for an examinee j .

$$\hat{\theta}_{j1} = E(\theta_{j1} | \mathbf{u}_j, \theta_{j2} \cdots \theta_{jk}) = \frac{1}{P_l} \int_{\theta_1} \theta_{j1} \left\{ \prod_{k=2}^s \int_{\theta_k} L_j(\theta_1, \theta_k) g(\theta_k) d\theta_k \right\} g(\theta_1) d\theta_1, \quad (4)$$

where P_l is the unconditional probability of observing response pattern $\mathbf{u}_j$ described in Equation 2, s is the number of group factors, and $L_j(\theta_1, \theta_k)$ is the likelihood function of the bifactor model for examinee j . Similarly, the posterior variance of $\hat{\theta}_{j1}$, which can be used to express the precision of the EAP estimator, is given by

$$V(\theta_{j1} | \mathbf{u}_j, \theta_{j2} \cdots \theta_{jk}) = \frac{1}{P_l} \int_{\theta_1} (\theta_{j1} - \hat{\theta}_{j1})^2 \left\{ \prod_{k=2}^s \int_{\theta_k} L_j(\theta_1, \theta_k) g(\theta_k) d\theta_k \right\} g(\theta_1) d\theta_1 \quad (5)$$

The EAP estimate of θ_k for examinee j is given by

$$\hat{\theta}_{jk} = E(\theta_{jk} | \mathbf{u}_j, \theta_{j1}) = \frac{1}{P_l} \int_{\theta_k} \theta_{jk} \int_{\theta_1} \left\{ L_j(\theta_1, \theta_k) \frac{\prod_{k=2}^s E_{jk}(\theta_1, \theta_k)}{E_{jk}(\theta_1, \theta_k)} g(\theta_1) d\theta_1 \right\} g(\theta_k) d\theta_k \quad (6)$$

where $E_{jk}(\theta_1, \theta_k) = \int_{\theta_k} L_j(\theta_1, \theta_k) g(\theta_k) d\theta_k$, and the corresponding posterior variance of $\hat{\theta}_{jk}$ is

$$V(\theta_{jk} | \mathbf{u}_j, \theta_{j1}) = \frac{1}{P_l} \int_{\theta_k} (\theta_{jk} - \hat{\theta}_{jk})^2 \int_{\theta_1} \left\{ L_j(\theta_1, \theta_k) \frac{\prod_{k=2}^s E_{jk}(\theta_1, \theta_k)}{E_{jk}(\theta_1, \theta_k)} g(\theta_1) d\theta_1 \right\} g(\theta_k) d\theta_k. \quad (7)$$

To evaluate the integrals, these integrations can be reasonably approximated using the Gauss-Hermite quadrature nodes and weights (see Stroud & Sechrest, 1996). θ s estimated by the EAP method have posterior variances after administration of each item. The OSE in EAP was computed by taking the square root of the posterior variance for each latent trait in Equations 5 and 7.

Item Selection for MCAT

After estimating current θ , next items should be selected by a specific item selection method in adaptive testing. Different item selection methods select different items for

examinees with equal θ . Therefore, the decision of item selection method should be considered in evaluating the accuracy and efficiency of the MCAT.

D-Optimality. Because this criterion maximizes the determinant in Equation 9, it is called D-optimality. In UCAT, items can be selected on the basis of item information. Likewise, in MCAT, the provisional trait estimate vector, $\hat{\theta}^{(n)}$ obtained after responding to the n th item, is used to evaluate the item information function (Lord, 1980):

$$I(\theta, u_i) = \frac{\left[\frac{\partial P_i(\hat{\theta}^{(n)})}{\partial \theta} \right]^2}{P_i(\hat{\theta}^{(n)}) Q_i(\hat{\theta}^{(n)})}, \quad (8)$$

where u_i is the candidate item response among items in the item bank, $P_i(\hat{\theta}^{(n)})$ is the item response function with candidate item i at $\hat{\theta}^{(n)}$, and $Q_i(\hat{\theta}^{(n)}) = 1 - P_i(\hat{\theta}^{(n)})$. The only difference from UCAT item selection is that MCAT item selection is based on the volume of the multivariate normal ellipsoid (Anderson, 1984). Similar to UCAT, if $\hat{\theta}^{(n)}$ is obtained from responding to the n th item, candidate item i is selected to maximize the determinant of information described as

$$\left| \mathbf{I}_{i|S_{n-1}}(\hat{\theta}^{(n)}, u_i) \right|, \quad (9)$$

where S_{n-1} is the set of administered items ($i_1, i_2, \dots, i_{n-1}$), and $\mathbf{I}_{i|S_{n-1}}$ implies that the information matrix associated with item i depends on both the characteristics of the candidate item i itself and the characteristics of the previously administered $n - 1$ items. The candidate item i , which maximizes the determinant of the information matrix $\mathbf{I}_{i|S_{n-1}}$ at $\hat{\theta}^{(n)}$ will provide the largest decrement in the size of the credibility region. The Bayesian item selection method (Segall, 1996) adjusts the maximum likelihood (ML) item selection method in Equation 9 for selecting candidate item i by maximizing the determinant of the posterior information matrix as

$$\left| \mathbf{I}_{i|S_{n-1}}(\hat{\theta}^{(n)}, u_i) + \Phi^{-1} \right|, \quad (10)$$

where Φ^{-1} is the inverse of the covariance matrix of the prior distribution of trait vector θ .

D_s-Optimality. Mulder and van der Linden (2009) stated that D_s-optimality (Silvey, 1980) reflects the optimal item selection for MCAT if the first ability of the ability vector θ is “intentional” and the last abilities are “nuisances.” In the bifactor model, the general factor measures “intentional” ability and group factors consider the “nuisance” abilities (Stucky, Thissen, & Edelen, 2013). Therefore, if θ_1 is assumed to be the only “intentional” ability, the appropriate weight vector is $\mathbf{A}^T = [1 \ 0 \ 0]$. Consequently, D_s-Optimality selects the item that minimizes the sampling variance of θ_1 as follows:

$$\min \left[\mathbf{A}^T \mathbf{I}_{i|S_{n-1}} \left(\hat{\boldsymbol{\theta}}^{(n)}, u_i \right)^{-1} \mathbf{A} \right]. \quad (11)$$

A Bayesian version of D_s -optimality was applied in this study by adding the inverse of a prior covariance matrix to Equation 11.

$$\min \left[\mathbf{A}^T \mathbf{I}_{i|S_{n-1}} \left(\hat{\boldsymbol{\theta}}^{(n)}, u_i \right)^{-1} \mathbf{A} + \boldsymbol{\Phi}^{-1} \right]. \quad (12)$$

Mulder and van der Linden (2009) showed that this criterion generally selects items that highly discriminate with respect to the “intentional” ability, θ_1 , except if the amount of information about the “nuisance” abilities is relatively low.

A-Optimality. The A-optimality method minimizes the sum of the asymptotic variances of the estimates, resulting in selection of the item that minimizes the traces of the inverse to the information matrix,

$$\text{trace} \left(\mathbf{I}_{i|S_{n-1}} \left(\hat{\boldsymbol{\theta}}^{(n)}, u_i \right)^{-1} \right). \quad (13)$$

Because A-optimality results in an item selection criterion that contains the determinant of the information matrix as an important factor, it is similar to the D-optimality method but is different from the D_s -optimality method (Mulder & van der Linden, 2009). A Bayesian version of A-optimality was applied in this study by adding the inverse of a prior covariance matrix to Equation 13:

$$\text{trace} \left(\mathbf{I}_{i|S_{n-1}} \left(\hat{\boldsymbol{\theta}}^{(n)}, u_i \right)^{-1} + \boldsymbol{\Phi}^{-1} \right). \quad (14)$$

E-Optimality. The criterion of E-optimality maximizes the smallest eigenvalues of the information matrix. The minimum of the eigenvalues of $\mathbf{I}_{i|S_{n-1}}$ is consistent across all item selection steps so the selection of the new item would not lead to any change of the current information matrix $\mathbf{I}_{S_{n-1}}$. According to the criterion of E-optimality, the candidate items contain no information about the θ estimates, and sampling variances of the θ estimators become equal to each other (Mulder & van der Linden, 2009).

Implementing the MCAT Algorithm

In the MCAT algorithm, θ estimation and item selection proceeded for all dimensions simultaneously. To compare four item selection and two θ estimation methods, the CAT was fixed length. Weiss and Gibbons (2007) showed that the mean number of items administered in CAT ranged from approximately 20 to 50 items per scale to recover each scale score with a correlation greater than .90 for all group factor scales. Therefore, MCAT in the present study terminated after 40 items were administered in the bifactor model with two group factors. The response pattern, current θ estimates,

and OSE for each examinee were saved after each item was administered. Because there was no available commercial software to implement the MCAT algorithm with the bifactor model, the MCAT algorithms were developed in the R language (R Development Core Team, 2012) by the author. To validate the program, θ estimates obtained by two estimation methods for the full-length MCAT algorithm were compared with those computed by “mirt” R package (Chalmers, 2012). The two sets of factor score matrices were identical to each other, rounded to two decimal places.

Evaluative Criteria

Four factors were examined: (1) three bifactor patterns, (2) four item selection methods, (3) two θ estimation methods, and (4) two MIRT models. The θ estimates obtained from administering a fixed number of items were used to evaluate the performance of the CAT. The fidelity, defined as the correlation between estimated θ and true θ (Weiss, 1982), was computed by Pearson product-moment correlations, $r(\theta_j, \hat{\theta}_j)$. The accuracy was evaluated by the RMSE, with lower values reflecting higher accuracy. The RMSE was computed as

$$RMSE(\theta_{jk}) = \sqrt{\frac{1}{N} \sum_{j=1}^N (\theta_{jk} - \hat{\theta}_{jk})^2}, \quad (15)$$

where j is an examinee, k is each factor, and N is the number of examinees. The efficiency of $\hat{\theta}$ s was evaluated by the OSE, with lower values reflecting greater efficiency. For MAP, the OSE in the MCAT was computed as

$$OSE(\hat{\theta}_j) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L}{\partial \theta_j^2}\right)}}. \quad (16)$$

The OSE of the EAP method in the MCAT was obtained by taking the square root of the posterior variances in Equations 5 and 7. These indices provided descriptive information about the recovery of θ for comparisons across the different cells in the research design. In the CAT literature, most studies have not considered replication if the overall performance of a CAT algorithm across θ levels is a major concern (e.g., Huang et al., 2012). Therefore, no replications were implemented in this study.

Results

Estimation Issues

The first operational problem for θ estimation methods in the MCAT with the bifactor model was to satisfy the convergence criterion of .001 because of difficulty

Table 2. Correlations Among $\hat{\theta}$ s for the Bifactor Model With Two Group Factors.

Model	Estimation	Bifactor low			Bifactor high			Bifactor-like		
		$r(G, g_1)$	$r(G, g_2)$	$r(g_1, g_2)$	$r(G, g_1)$	$r(G, g_2)$	$r(g_1, g_2)$	$r(G, g_1)$	$r(G, g_2)$	$r(g_1, g_2)$
2-PL	MAP	.086	.084	-.322	.089	.086	-.321	.092	-.031	-.107
	EAP	.092	.092	-.362	.091	.098	-.359	.093	-.074	-.131
3-PL	MAP	.092	.095	-.330	.096	.097	-.283	.098	-.047	-.181
	EAP	.093	.097	-.346	.097	.099	-.303	.099	-.075	-.208

Note. MAP = maximum a posteriori; EAP = expected a posteriori. $r(G, g_1)$ is correlation between general factor and the first group factor, $r(G, g_2)$ is correlation between general factor and the second group factor, and $r(g_1, g_2)$ is correlation between the first group factor and the second group factor.

meeting simultaneously the convergence criterion of .001 for all latent traits. Therefore, this study used θ estimation algorithms with a maximum of 10 iterations instead of the convergence criterion of .001. To ensure convergence of 1,000 simulees' estimates across each condition, the convergence criterion of each factor was investigated by an R program. The average convergence criteria for each latent trait of 1,000 simulees were less than .001 for all latent traits when the maximum number of iterations was set to 10. The second computational issue was to find MLE estimates in the MCAT algorithm with the bifactor model. The MLE estimation method in the bifactor model has a computational issue in that the Hessian matrix becomes singular during updating θ estimates. Because a singular matrix cannot be inverted, iterations were immediately terminated. The Hessian matrix always became singular if the bifactor model was applied to MCAT to update the θ estimates even if the response sets were manipulated as mixed responses with 0s and 1s. The MLE method in the "mirt" R package (Chalmers, 2012) also did not estimate θ vectors under the bifactor model. Therefore, this study excluded the MLE method for MCAT with the bifactor model.

The bifactor model in this study was constructed so that uncorrelated group factors were independent of the general factor (Holzinger & Swineford, 1937). According to the model assumption, θ s should be independent of each other. Table 2 shows correlations among θ s for three group factor pattern designs using two θ estimation methods. Results showed that the MCAT with the bifactor model relatively satisfied the independence assumption of θ s between the general factor and group factors, whereas the group factors were negatively correlated with each other across all conditions. Equation 2 indicates that primary and secondary factors are distributed independently in the examinee population when items are selected in MCAT so that the general factor was uncorrelated with the group factors. The two group factors, however, were negatively correlated with each other because the MCAT with the bifactor model proceeds separately for each group factor. The correlations between the general factor and group factors were not affected by the increase of group factor loadings, whereas the correlations between two group factors for 3PL bifactor were closer to zero when

Table 3. $r(\theta, \hat{\theta})$ Across Four Item Selection Algorithms for Three Bifactor Pattern Designs.

Model	Item selection	Estimation	Bifactor low			Bifactor high			Bifactor-like		
			G	g_1	g_2	G	g_1	g_2	G	g_1	g_2
2PL	D-Optimality	MAP	.975	.774	.817	.927	.906	.927	.961	.850	.860
		EAP	.973	.770	.812	.923	.892	.912	.959	.844	.847
	A-Optimality	MAP	.974	.785	.811	.917	.909	.919	.959	.863	.888
		EAP	.971	.782	.809	.913	.906	.909	.957	.862	.866
	D_s -Optimality	MAP	.979	.695	.798	.958	.749	.811	.975	.788	.793
		EAP	.982	.693	.780	.974	.746	.803	.978	.783	.789
	E-Optimality	MAP	.967	.656	.789	.892	.772	.751	.972	.724	.733
		EAP	.944	.586	.444	.734	.695	.484	.821	.460	.408
3PL	D-Optimality	MAP	.965	.711	.579	.922	.812	.838	.959	.757	.742
		EAP	.969	.661	.547	.920	.805	.812	.921	.740	.723
	A-Optimality	MAP	.964	.714	.572	.912	.822	.829	.956	.760	.740
		EAP	.954	.678	.538	.910	.814	.822	.931	.741	.727
	D_s -Optimality	MAP	.976	.669	.560	.949	.736	.746	.972	.621	.596
		EAP	.978	.642	.555	.952	.681	.740	.977	.620	.533
	E-Optimality	MAP	.914	.604	.503	.820	.715	.712	.797	.577	.522
		EAP	.884	.601	.333	.818	.567	.529	.737	.534	.444

Note. MAP = maximum a posteriori; EAP = expected a posteriori. G is the general factor, g_1 is the first group factor, and g_2 is the second group factor.

the average group factor pattern was higher. The EAP estimation method showed that correlations between the general factor and group factors were slightly higher, and correlations between two group factors were moderately higher than those from the MAP estimation method. Equations 4 and 6 indicate that primary factor scores were estimated first, and then secondary factor scores were estimated in the EAP estimation method so that correlations between the two group factors from the EAP method were higher than those from the MAP method. The bifactor-like pattern showed that the general factor was uncorrelated with the group factor, and correlations between group factors were closer to zero than those from the two typical bifactor pattern designs.

Fidelity, Accuracy, and Efficiency

Tables 3 and 4 summarize $r(\theta, \hat{\theta})$ and RMSE, respectively, of 1,000 simulees for the general factor and the group factors across all conditions. Results across three bifactor pattern designs showed that D_s -optimality item selection provided higher $r(\theta, \hat{\theta})$ and lower RMSE than other item selection methods for the general factor under two MIRT models, whereas it indicates lower $r(\theta, \hat{\theta})$ and higher RMSE than other item selection methods for group factors under two MIRT models. These results suggest that D_s -optimality is the best of the four item selection methods studied if focus is on measuring the general factor. Because it is common to focus on the general

Table 4. RMSEs Across Four Item Selection Algorithms for Three Bifactor Pattern Designs

Model	Item selection	Estimation	Bifactor low			Bifactor high			Bifactor-like		
			G	g_1	g_2	G	g_1	g_2	G	g_1	g_2
2PL	D-Optimality	MAP	.234	.607	.601	.247	.471	.391	.241	.557	.509
		EAP	.244	.616	.605	.256	.498	.408	.254	.568	.519
	A-Optimality	MAP	.227	.604	.602	.274	.459	.451	.242	.541	.524
		EAP	.238	.609	.607	.266	.467	.464	.269	.559	.538
	D_s -Optimality	MAP	.215	.646	.642	.216	.525	.501	.157	.644	.620
		EAP	.212	.648	.641	.214	.538	.521	.148	.673	.652
	E-Optimality	MAP	.256	.653	.651	.296	.526	.565	.529	.708	.821
		EAP	.335	.791	.677	.457	.578	.788	.549	.715	.866
3PL	D-Optimality	MAP	.303	.701	.729	.362	.715	.638	.283	.667	.735
		EAP	.292	.754	.744	.374	.690	.653	.321	.865	.747
	A Optimality	MAP	.305	.677	.735	.417	.705	.647	.261	.626	.772
		EAP	.345	.705	.753	.392	.651	.665	.294	.833	.767
	D_s -Optimality	MAP	.275	.720	.783	.341	.751	.691	.235	.885	.897
		EAP	.273	.763	.787	.340	.780	.723	.233	.846	.939
	E-Optimality	MAP	.313	.759	.795	.422	.763	.781	.531	.903	1.108
		EAP	.411	.769	.868	.630	.841	.865	.602	.929	1.111

Note. MAP = maximum a posteriori; EAP = expected a posteriori; RMSE = root mean square error. G is the general factor; g_1 is the first group factor; and g_2 is the second group factor.

dimension and then consider the subscale dimensions in many psychometric analyses, D_s -optimality is appropriate for MCAT with the bifactor model if the general factor is “intentional,” and group factors are considered “nuisance” variables. For the group factor scales, D-optimality or A-optimality resulted in the most accurate estimates across the three bifactor pattern designs under the two MIRT models. This result was expected because D- or A-optimality select items that highly discriminate θ with respect to group factors. E-optimality as an item selection method provided the lowest $r(\theta, \hat{\theta})$ and the highest RMSE across all conditions. In E-optimality, the contribution of an item with equal discrimination parameters to the test information vanishes if the sampling variances of the θ estimators become equal to each other (Mulder & van der Linden, 2009). Therefore, E-optimality did not select items that differentiate θ s in MCAT with the bifactor model.

The precision of $\hat{\theta}$ in MCAT varied according to the group factor pattern designs. The bifactor models with low group factor discriminations had higher $r(\theta, \hat{\theta})$ and lower RMSE than the other two designs with respect to the general factor, and lower $r(\theta, \hat{\theta})$ and larger RMSE than the other two designs with respect to the group factors. The bifactor models with high group factor patterns had lower $r(\theta, \hat{\theta})$ and higher RMSE than the other two designs with respect to the general factor, and higher $r(\theta, \hat{\theta})$ and lower RMSE than the other two designs with respect to the group factors. In sum, the recovery of group factor scores was improved as the average group factor discrimination parameters became high, whereas the recovery of general factor scores was exacerbated as the average group factor discrimination parameters became high.

Table 5. OSEs of Estimates Across Four Item Selection Algorithms for Three Bifactor Pattern Designs.

Model	Item selection	Estimation	Bifactor low			Bifactor high			Bifactor-like		
			G	g_1	g_2	G	g_1	g_2	G	g_1	g_2
2PL	D-Optimality	MAP	.171	.640	.446	.203	.443	.445	.182	.519	.520
		EAP	.174	.655	.452	.221	.449	.457	.184	.521	.522
	A-Optimality	MAP	.185	.625	.446	.203	.432	.447	.191	.511	.512
		EAP	.192	.641	.479	.231	.448	.462	.193	.523	.523
	D_s -Optimality	MAP	.170	.667	.532	.203	.463	.477	.185	.603	.605
		EAP	.162	.729	.540	.191	.567	.542	.173	.617	.619
	E-Optimality	MAP	.206	.675	.727	.426	.631	.700	.443	.649	.750
		EAP	.321	.829	.797	.429	.642	.777	.446	.677	.788
3PL	D-Optimality	MAP	.211	.717	.508	.253	.528	.543	.232	.616	.621
		EAP	.215	.745	.529	.254	.529	.550	.233	.624	.627
	A-Optimality	MAP	.231	.691	.523	.253	.510	.550	.232	.601	.614
		EAP	.245	.720	.539	.256	.518	.553	.247	.617	.636
	D_s -Optimality	MAP	.210	.706	.640	.253	.550	.551	.231	.727	.746
		EAP	.195	.710	.654	.249	.558	.559	.198	.737	.754
	E-Optimality	MAP	.255	.711	.781	.467	.673	.810	.578	.774	.850
		EAP	.394	.824	.897	.470	.689	.867	.584	.780	.897

Note. OSE = observed standard error; MAP = maximum a posteriori; EAP = expected a posteriori. G is the general factor, g_1 is the first group factor, and g_2 is the second group factor.

MAP estimation showed higher $r(\theta, \hat{\theta})$ and lower RMSE values than EAP estimation for all conditions except general factors using D_s -optimality but differences in $r(\theta, \hat{\theta})$ between the two θ estimation methods were minimal across all conditions. Estimation method did not much affect measurement fidelity in this study. This result confirmed that MAP performs similarly or better than EAP in general MCAT algorithms (Yao, 2013). An interesting finding in this study was that EAP estimation performed better with the D_s -optimality item selection method in estimating the general factor scores. Therefore, if interest is in general factor scores in MCAT with the bifactor model, EAP estimation with D_s -optimality item selection will be the best design to estimate general scores. Comparing the two MIRT models, the 2PL the bifactor model provided higher $r(\theta, \hat{\theta})$ and lower RMSE than the 3PL bifactor model across all conditions for both general and group factors.

Table 5 displays the average OSEs of the θ estimates using four item selection algorithms for three bifactor designs. Similar to the results of $r(\theta, \hat{\theta})$ and RMSE, the D_s -optimality provided the smallest OSEs among the four item selection methods for the general factor, whereas the OSEs from D- or A-optimality were the smallest for the group factors. There were OSE differences among the two θ estimation methods across the four item selection algorithms: EAP estimation with D_s -optimality provided the smallest OSE for the general factor, whereas EAP estimation showed slightly larger OSEs in the general and group factors in most conditions except for

the general factors with D_s -optimality. In the EAP computation, the general dimension is considered first and then group factors are considered to estimate the probability of a correct response in the bifactor model. Based on this sequential computation, EAP estimation performed well with the D_s -optimality item selection method for the general factor scores in MCAT with the bifactor model.

Similar to the results of correlation and RMSE, OSEs for the group factors in the bifactor pattern with high group factor discrimination parameters were smaller than the other bifactor pattern designs across all conditions, whereas the general factor OSEs in the bifactor design with low group factor discrimination were smaller than the other bifactor pattern designs across all conditions. If the average group factor discrimination parameters increased, OSEs of the group factors were decreased. Because higher item information considers the decline of OSEs, the increase in the group factor discrimination parameters provided low OSEs for group factors in MCAT with the bifactor model. In the model comparison, the 2PL bifactor model provided lower OSEs than the 3PL bifactor model across all conditions.

Discussion and Conclusions

This study applied the bifactor model to MCAT under varied bifactor pattern designs using two MIRT models, three θ estimation methods, and four MCAT item selection methods. Results showed that MLE cannot be applied within the bifactor model because of singularity of the information matrix. Therefore, only MAP and EAP θ estimation methods were analyzed. By examining the correlations among estimated θ s, this study demonstrated that the MCAT algorithm satisfied the independence of the general factor and the group factors. An interesting finding in this study was that D_s -optimality item selection worked well with EAP estimation for the general factor scores. For the group factors, however, D- or A-optimality item selection improved both the accuracy and efficiency of θ estimates. Therefore, if researchers are primarily interested in general factor θ s, D_s -optimality and the EAP estimation method are recommended, whereas if interest is primarily in θ s for the group factors, D- or A-optimality item selection methods and MAP estimation method would be appropriate. Using E-optimality for item selection in the bifactor MCAT resulted in low accuracy of scores, and is not recommended.

High group factor discrimination parameters contributed to higher precision and better efficiency of only group factor θ estimation across all conditions. When the average group factor discrimination parameters were high, precision of the estimated θ s for the group factors improved, whereas precision of the estimated θ s for the general factor was slightly lowered. In terms of efficacy of the CAT algorithm, when the average group factor discrimination parameters increased, OSE for the general factor scores slightly increased, whereas the OSE for the group factors decreased. These results imply that increasing group factor discrimination parameters improves the accuracy and efficacy of group factor scores, and it reduces slightly the accuracy and efficiency of general factor scores in bifactor MCAT. This study suggests that MCAT

with a bifactor model is especially appropriate for the bifactor model with a dominant general factor. Results showed that EAP estimation was less efficient than MAP estimation except for the general factor using D_s -optimality. The OSE (variance-covariance matrix) of MAP estimation in Equation 14 is the inverse of the information matrix so that MAP algorithm usually reduces the full item information matrix to a single value as a composite function to select the following item that optimizes item information. Therefore, MAP incorporates the necessary content information so that it does guarantee minimum OSEs reflecting global objective information (Luecht, 1996). However, the EAP approach estimates general factor θ s first in Equation 4 and then estimates group factor θ s in Equation 6, thereby estimating general and group factors θ s separately. This sequential process will not guarantee the minimum OSEs reflecting global objective information. Consequently, OSEs of EAP θ estimates were higher than those from MAP because the EAP method ignored more content information from the full information matrix.

Although development of MIRT began many years ago, there are few practical applications of MIRT to CAT because of its complexity. The advantages of MCAT with the bifactor model are that (1) it has higher fidelity and accuracy with respect to a k -dimensional vector of traits for each examinee as compared with using successive unidimensional CATs and (2) it satisfies independence of the general factor θ estimates and the group factor θ estimates (the general factor score can be interpreted as the unidimensional trait in UCAT). In the future, therefore, continued research on MCAT with the bifactor model is warranted and the analytic techniques in this study should help researchers who would implement MCAT with the bifactor model find the systematic and optimal item selection and estimation methods that are appropriate under certain circumstances for practical application.

However, some obvious issues are raised by this study for future research. This study used the bifactor model to implement MCAT as one specific model among many possible multidimensional models. By considering the effect of multidimensionality on IRT item parameter estimates, the bifactor model estimates a substantively common trait as well as capturing multiple constructs (Reise et al., 2007). However, the bifactor model is not versatile for all psychological data. Gibbons et al. (2008) mentioned some limitations of the bifactor model. First, the bifactor model specification relies on prior information to indicate the relationships between items and factors. Second, a primary (i.e., general) dimension is assumed to exist. If the test items show a strong simple factor structure, the bifactor model will not be useful. Under these circumstances, it would be appropriate to use a unidimensional model for each factor structure rather than the bifactor model. This model comparison is empirically testable by comparing the fit of the bifactor model with corresponding unidimensional models and unrestricted multidimensional models by using TESTFACT (Wood et al., 2003) or the “mirt” R package (Chalmers, 2012). Third, the bifactor model requires each item to load on a primary dimension and on no more than one subdomain. If items are related to multiple subdomains, they will not be appropriate for the bifactor model. Therefore, the bifactor model can be adapted

to MCAT only if the researcher has strong theoretical beliefs concerning the structure of a domain or has evidence concerning the actual dimensional structure of a test, and after demonstrating that the bifactor model fits better than alternative models.

In this study, only a small subset of bifactor structures was examined to investigate the quality of MCAT with the bifactor model. In this regard, this study has several limitations that could be of interest in future studies. First, this study generated the data using specific instances of the bifactor structure. This means that the mean and standard deviation of group factor discrimination parameters for each group factor were applied equally to generate group factor discrimination parameters for each item. Second, this study used only three bifactor loading conditions (with low group factor loadings, high group factor loadings, and bifactor-like loadings). Future studies should investigate whether the same conclusions can be arrived if the conditions are further varied. Furthermore, additional multidimensional item selection methods can be considered, such as maximizing the Kullback-Leibler information (Veldkamp & van der Linden, 2002). Yao (2013) recommended the Kullback-Leibler method for variable-length MCAT. Possible directions could be usage of the complete information matrix to capture more information rather than applying the determinant alone or a simple algorithm that reduces the computational load. A new item selection method comprising the bifactor model needs to be developed and compared with usual multidimensional item selection methods in future research. Additionally, there are also other potential research issues in MCAT with the bifactor model, such as how to select the first item and how to control item exposure. MCAT with the bifactor model could resolve the content-balance issue in practical testing. Because each item in the bifactor model loads on only one specific group factor, MCAT with the bifactor model alternated items that loaded on each group factor, which functioned as content balancing. MCAT with the bifactor model relatively administered an equal number of items with respect to each of the group factor scales, which would result in content-balanced θ estimates based on various mixtures of group factor scales, which is an additional advantage of MCAT with the bifactor model.

Finally, researchers should consider some computational issues when MCAT with the bifactor is applied to practical data. First, MLE did not operate in the MCAT with the bifactor model. As mentioned above, the Hessian matrix ($p \times p$ symmetric matrix) of second derivatives (the negative information matrix) evaluated at $\theta^{(n)}$ should be inverted to update previous estimates during iterations. However, the Hessian matrix could not be inverted because it was always a singular matrix. Even if the MLE procedures in MCAT took the second derivative of the likelihood function with additional prior multivariate standard normal distributions, a Hessian matrix could not be inverted. Second, EAP has a computational burden to recover θ , even if a small number of quadrature points were applied in EAP estimation. For the condition involving a bank of 400 items and 1,000 examinees in MCAT with two group factors, the average time required to estimate θ using a convergence criterion of .001 on a computer with a 3.2 GHz processor and 2 GB of memory was

approximately 25 minutes for MAP but was more than 4 hours for EAP using 15 quadrature points for the general factor and each group factor. The bifactor model has an advantage to reduce the s -dimensional integral into a two-dimensional integral as a computational efficiency. Nevertheless, there is still significant computational burden for EAP in the MCAT algorithm.

Declaration of Conflicting Interests

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

Funding

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

References

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.
- Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: John Wiley.
- Ansley, R. A., & Forsyth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.
- Baker, F. B. (1992). *Item response theory parameter estimation techniques*. New York, NY: Marcel Dekker.
- Bloxom, B. M., & Vale, C. D. (1987, June). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavior Research, 36*, 111-150.
- Cai, L., Yang, J., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221-248.
- Cassano, G. B., Michelini, S., Shear, M. K., Coli, E., Maser, J. D., & Frank, E. (1997). The panic-agoraphobic spectrum: A descriptive approach to the assessment and treatment of subtle symptoms. *American Journal of Psychiatry, 154*, 27-38.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. Retrieved from <http://www.jstatsoft.org/v48/i06>
- Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in a stratified computerized adaptive testing. *Applied Psychological Measurement, 27*, 262-274.

- Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41, 189-224.
- Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R. Professional manual*. Odessa, FL: Psychological Assessment Resources.
- DeMars, C. E. (2007). "Guessing" parameter estimates for multidimensional item response theory models. *Educational and Psychological Measurement*, 67, 433-446.
- Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, 13, 373-390.
- Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities. Lecture notes in statistics* (Vol. 195). Heidelberg, Germany: Springer-Verlag.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., & . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31, 4-19.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika*, 57, 423-436.
- Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., & . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, 59(4), 49-58.
- Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97-121). Washington, DC: American Psychological Association.
- Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 134, 486-501.
- Holzinger, K. J., & Swineford, F. (1937). The bifactor method. *Psychometrika*, 2, 41-54.
- Horn, J. L. (1986). Intellectual ability concepts. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 3, pp. 35-78). Hillsdale, NJ: Lawrence Erlbaum.
- Huang, H.-Y., Chen, P.-H., & Wang, W.-C. (2012). Computerized adaptive testing using a class of high-order item response theory models. *Applied Psychological Measurement*, 36, 689-706.
- Jennrich, R., & Bentler, P. (2012). Exploratory bi-factor analysis: the oblique case. *Psychometrika*, 77, 442-454.
- Jöreskog, K. G., & Sörbom, D. (1995). LISREL 8 user's reference guide [Computer software]. Chicago, IL: Scientific Software.
- Kingsbury, G. G., & Weiss, D. J. (1980). *An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests* (Research Report 80-5). Minneapolis: University of Minnesota.
- Kingsbury, G. G., & Weiss, D. J. (1983). *A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure*. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 257-283). New York, NY: Academic Press.
- Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement*, 29, 3-25.
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389-404.
- Moreno, K. E., & Segall, D. O. (1992). CAT-ASVAB precision. *Proceedings of the 34th Annual Conference of the Military Testing Association*, 1, 22-26.
- Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, 74, 273-296.
- Muthén, L. K., & Muthén, B. O. (2004). Mplus user's guide (Version 3) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.
- R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for statistical Computing.
- Reckase, M. D. (1974). An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods and Instrumentation*, 6, 208-212.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92, 544-559.
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16, 19-31.
- Revelle, W. (2015). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from <http://cran.r-project.org/package=psych>
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.
- Silvey, S. D. (1980). *Optimal design*. London, England: Chapman & Hall.
- Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice Hall.
- Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*, 37, 41-57.
- Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait* (Unpublished doctoral dissertation). Columbia University, New York, NY.
- van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63, 201-216.
- Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67, 575-588.
- Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 28, 295-316.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.
- Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53, 774-789.
- Weiss, D. J., & Gibbons, R. D. (2007, June). *Computerized adaptive testing with the bifactor model*. Paper presented at the New CAT Models session at the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved from <http://publicdocs.iacat.org/cat2010/cat07weiss&gibbons.pdf>

- Wolff, H., & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. *Behavioral Research Methods*, 37, 48-58.
- Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). TESTFACT 4 [Computer software]. Lincolnwood, IL: Scientific Software.
- Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedure with different stopping rules. *Applied Psychological Measurement*, 37, 3-23.