Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 19, Number 2, May 2014 ISSN 1531-7714

Impact of Violation of the Missing-at-Random Assumption on Full-Information Maximum Likelihood Method in Multidimensional Adaptive Testing

Kyung T. Han & Fanmin Guo
Graduate Management Admission Council®

The full-information maximum likelihood (FIML) method makes it possible to estimate and analyze structural equation models (SEM) even when data are partially missing, enabling incomplete data to contribute to model estimation. The cornerstone of FIML is the missing-at-random (MAR) assumption. In (unidimensional) computerized adaptive testing (CAT), unselected items (i.e., responses that are not observed) remain at random even though selected items (i.e., responses that are observed) have been associated with a test taker's latent trait that is being measured. In multidimensional adaptive testing (MAT), however, the missingness in the response data partially depends on the unobserved data because items are selected based on various types of information including the covariance among latent traits. This eventually may lead to violations of MAR. This study aimed to evaluate the potential impact such a violation of MAR in MAT could have on FIML estimation performance. The results showed an increase in estimation errors in item parameter estimation when the MAT response data were used, and differences in the level of the impact depending on how items loaded on multiple latent traits.

Although the technical and practical frameworks of factor analysis (FA) and item response theory (IRT) were developed independently from one another, the literature reveals an obvious connection between FA and IRT such that one approach essentially can yield results equivalent to those from the other approach under various conditions (Takane & Leeuw, 1987; Reise, Widaman, & Pugh, 1993; Kamata & Bauer, 2008). Just as the IRT framework and its initial applications, which are based mainly on a unidimensional latent structure (Lord & Novick, 1968), were extended for a variety of multidimensional latent structures, so too have the relations between multidimensional IRT (MIRT) and FA (particularly, the confirmatory factor analysis (CFA) and the structural equation modeling (SEM)) been revisited and studied (McDonald, 2000; Reckase, 2009; Osteen, 2010).

Efforts to incorporate MIRT into computerized adaptive testing (CAT) have made significant progress as well (Segall, 1996, 2000; Reckase, 2009). In order to analyze response data from multidimensional adaptive testing (MAT) using SEM, however, one must first address technical obstacles related to the uniqueness of the CAT response data—the extreme level of sparseness of the data matrix and its *missing* mechanism, which does not strictly meet the *missing-at-random (MAR)* condition. The purpose of this study is to evaluate the performance of the full-information maximum likelihood (FIML) method when using response data sets from MAT.

Relations Between IRT and CFA

One of the most common IRT models for dichotomous responses with a single latent trait is the two-parameter logistic (2PL) model, which can be expressed as

$$P\left(U_{ij}=1\middle|\theta_j,a_i,b_i\right)=\frac{1}{1+\exp(-a_i(\theta_j-b_i))} \tag{1}$$

where θ_j is the person parameter describing the characteristics on the relevant latent trait of examinee *j*, and a_i and b_i are the item parameters describing the discrimination and difficulty of item *i*. When there is more than one latent trait of interests, the 2PL model (Equation 1) can be generalized as

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta}_{\mathbf{j}},\mathbf{a}_{\mathbf{i}},b_i\right)=\frac{1}{1+\exp(-\mathbf{a}_{\mathbf{i}}^{'}(\boldsymbol{\theta}_{\mathbf{j}}-b_i\mathbf{1}))}, \quad (2)$$

where $\mathbf{a}_{\mathbf{i}}^{'}$ is a 1 × H vector of discrimination parameters for each relevant trait (with H being the number of latent traits), $\boldsymbol{\theta}_{\mathbf{j}}$ is a vector of person parameters for each H trait, and $\mathbf{1}$ is a H × 1 vector of 1's (the symbols in bold represent vectors). In Equation 2, the parameters of the exponent, $\mathbf{a}_{\mathbf{i}}^{'}\left(\boldsymbol{\theta}_{\mathbf{j}}-b_i\mathbf{1}\right)$, can be reparameterized as $\boldsymbol{\lambda}_{\mathbf{i}}^{'}\boldsymbol{\theta}_{\mathbf{j}}+\nu_i$, where $\boldsymbol{\lambda}_{\mathbf{i}}^{'}=\mathbf{a}_{\mathbf{i}}^{'}$ and $\nu_i=\mathbf{a}_{\mathbf{i}}^{'}b_i\mathbf{1}$. In $\boldsymbol{\lambda}_{\mathbf{i}}^{'}\boldsymbol{\theta}_{\mathbf{j}}+\nu_i$, $\boldsymbol{\nu}$ is often called the intercept and $\boldsymbol{\lambda}$ is called the set of slopes. With the reparameterization, it becomes clear that the exponent is essentially equivalent to a common factor analytical model (Christoffersson, 1975), which often is expressed as

$$y_{ij}^{*}=\nu_i+\boldsymbol{\lambda}_{\mathbf{i}}^{'}\boldsymbol{\xi}_{\boldsymbol{j}}+\varepsilon_i, \quad (3)$$

where y_{ij}^{*} is a latent response variable, $\boldsymbol{\xi}$ is a vector of factor scores of person *j* on each latent trait (which can be replaced by $\boldsymbol{\theta}_{\mathbf{j}}$), and, ε_i is the residual, which typically is assumed to be normally distributed. In cases with binary variables such as the multidimensional 2PL model a threshold model is added, where the observed binary response is

$$y_{ij}=\begin{cases}1 \text{ if } y_{ij}^{*}\geq\tau_i\\0 \text{ if } y_{ij}^{*}<\tau_i\end{cases}. \quad (4)$$

In practice, one typically deals only with the threshold, τ_i, assuming the intercept, ν_i, in Equation 3 to be zero (Takane & Leeuw, 1987; Kamata & Bauer, 2008).

Several studies applied, examined, and compared the CFA frameworks and methods to IRT-based methods. Takane and Leeuw (1987) analytically explained the equivalent relation between the FA and (unidimensional) IRT approaches, and several other studies including Reise et al. (1993) empirically showed the similarities between the two approaches using real data. A more direct comparison can be made between MIRT and FA approaches; in fact, earlier MIRT models such as those that Bock and Aitken (1981), Samejima (1974), and McDonald (1967) proposed clearly showed that the MIRT and FA share virtually identical mathematical models (Reckase, 2009).

CAT and Missing Mechanism

With the emergence of IRT, which enables tests to be analyzed and constructed at the item level, and the help of modern computers that are powerful enough to administer tests adaptively on the fly, computerized adaptive testing (CAT) has quickly become one of the most popular modes of testing. In CAT, test items that are expected to exhibit the highest information (or are expected to have the most relevant difficulty level) for each individual examinee are selected and administered adaptively based on examinee's performance on previously administered items (Lord, 1980). As a result, CAT usually exhibits measurement efficiency that exceeds that of tests not adaptively administered: equivalent or higher measurement quality with fewer test items (Weiss, 1974, 1982). Computerized adaptive testing for multidimensional cases (e.g., multidimensional adaptive testing or MAT) has developed naturally as the unidimensional IRT was extended to the multidimensional IRT (Reckase, 2009; Segall, 1996, 2000; Veldkamp & van der Linden, 2002).

With CAT/MAT, test developers always pretest, precalibrate, and preanalyze operational test items before adding them to an operational item bank, and typically use examinees' response data from operational administrations only for scoring. It is expected that response data from operational CAT administrations will contain useful information for continuous quality control of CAT programs, used, for example, to monitor for item parameter drifts by recalibrating items and reevaluating the latent structures using SEM. Such applications, however, have not yet been extensively studied. Response data from adaptive testing have not been used much with SEM analyses largely because of the unique *missingness* in the response matrix of CAT. In CAT, each examinee responds only to a fraction of the test items contained in the entire item pool. This makes the full response matrix (an *n* × *m* matrix with *n* being the total number of examinees and *m* being the total number of items in the item pool) very sparse. In high-stakes CAT programs, the item exposure rate usually is controlled to be minimal (often smaller than 0.1 to 0.2) for test security purposes, which makes the full response matrix extremely sparse. For a response matrix with such an extreme level of sparseness, most traditional methods for dealing with missingness of data become impractical, for example, the listwise deletion and the pairwise deletion for old FA approaches based on the least square method and its variations.

The emergence of the full information maximum likelihood (FIML) method—simply known as the maximum likelihood (ML) method—completely changed the way we deal with missing data because it does not require a complete response matrix with no missing data (Bartholomew, 1980; Enders & Bandalos, 2001; Graham, 2009; Schafer & Graham, 2002). Several estimators—the marginal maximum likelihood (MML) method (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988) and the expectation-maximization (EM) algorithm (Little & Rubin, 1987; Schafer, 1997)—were developed and used recently in several widely-used SEM software programs including *Mplus* (Muthén & Muthén, 2010), LISREL (Jöreskog & Sörbom, 2006), and Amos (Arbuckle, 2006). The FIML methods require a less restrictive "missing at random" (MAR) assumption, where the missingness depends on observed data but not on unobserved data. They also are known to result in unbiased estimates under both MAR and "missing completely at random" (MCAR) scenarios, where the missingness depends neither on observed nor unobserved data (Rubin, 1976; Enders & Bandalos, 2001; Graham, 2009).

The missingness of the response matrix in (unidimensional) CAT was often seen as satisfying MAR because the item selection process depends on examinees' observed performance on previous items (i.e., observed data) not on examinees' performance on unadministered items (i.e., unobserved data). With the ignorability by satisfying MAR, items and examinees can be calibrated using the MML method based on CAT data unless the CAT administration is extremely optimal at a true latent trait (Glas, 1988, 2010; Han, Guo, Talento-Miller, & Rudner, 2011).

In MAT, the item selection process considers more than one latent trait at a time. Unless the MAT is based on a completely noncompensatory MIRT model, or there is zero covariance among latent traits, or the item selection algorithm is focused solely on a single factor (e.g., the general factor in the bi-factor model approach), one can assume that an examinee's proficiency on one trait is related to other traits. This piece of information—the covariance matrix of latent traits—weighs heavily in MAT, and, as a result, an examinee's observed performance on one trait can affect the item selection process for items measuring other traits in MAT. In other words, the missingness of the MAT response data cannot be guaranteed to uphold the MAR assumption and it may be more appropriate to consider it as a "missing not at random" (MNAR) case, where the missingness depends on both observed and unobserved data.

The object of this study, then, is to evaluate the potential impact that violations of MAR in MAT may have on model estimation with SEM using the FIML (more specifically, MML) method.

Method

MAT Simulations

Our study involved conducting a series of MAT simulations. Three hundred test items were generated based on the multidimensional compensatory 2PL (MC2PL) model, in which the exponent of Equation 2 was reparameterized to $-\mathbf{a}_\mathbf{i}^{'}\boldsymbol{\theta}_\mathbf{j} + d_i$. The *d*-parameters (i.e., threshold parameters) were drawn from a normal distribution, and the actual sample mean was –0.313 with a standard deviation (SD) of 1.032. The test was designed to measure two latent traits—F1 and F2. The 300 items were classified into one of five groups by the *a*-parameter values (i.e., factor loadings). Group 1 items were loaded only on a single factor (either F1 or F2), and Group 2 items were loaded primarily on one factor and slighted on the other factors. Group 3 items were loaded and moderated on both factors but slightly more on one factor than the other. Group 4 items were loaded equally on both factors. The final group of items—Group 0—was loaded only on a single factor (similar to the Group 1 items) and fixed to be loaded only on the corresponding factor during model estimation. The *a*-parameter values are reported in Table 1. The overall latent structure of the items in the item pool is shown in Figure 1.

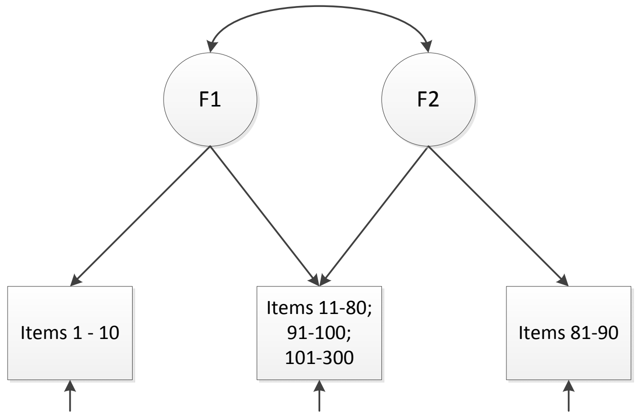

Figure 1. The structure of the test with two latent traits.

Ten thousand test simulees with two latent traits were generated. The person parameters for the first trait (θ_1) were sampled from a standard normal distribution; the actual sample mean was 0.006 with the variance of 0.998. The person parameters for the second trait (θ_2) were generated to correlate with θ_1 at the correlation coefficient of 0.710; the actual sample mean of θ_2 was 0.000 with the variance of 1.986. The sample covariance between θ_1 and θ_2 was 1.000.

In the simulations, 30 items were adaptively administered to each simulee. Two different item selection conditions were studied. In the first MAT condition, the items were selected based on the *maximized determinant of posterior information* (MDPI)

Table 1. *List of Item Parameters and Groups*

Item ID	$\mathbf{a}_{(1)}$	$\mathbf{a}_{(2)}$	Group^{+}	Item ID	$\mathbf{a}_{(1)}$	$\mathbf{a}_{(2)}$	Group^{+}
1,101,201	2.000	0.000	0 (for 1); 1(for rest)	51,151,251	0.900	1.100	3
2,102,202	2.000	0.000	0 (for 2); 1(for rest)	52,152,252	0.900	1.100	3
3,103,203	2.000	0.000	0 (for 3); 1(for rest)	53,153,253	0.900	1.100	3
4,104,204	2.000	0.000	0 (for 4); 1(for rest)	54,154,254	0.900	1.100	3
5,105,205	2.000	0.000	0 (for 5); 1(for rest)	55,155,255	0.900	1.100	3
6,106,206	2.000	0.000	0 (for 6); 1(for rest)	56,156,256	0.900	1.100	3
7,107,207	2.000	0.000	0 (for 7); 1(for rest)	57,157,257	0.900	1.100	3
8,108,208	2.000	0.000	0 (for 8); 1(for rest)	58,158,258	0.900	1.100	3
9,109,209	2.000	0.000	0 (for 9); 1(for rest)	59,159,259	0.900	1.100	3
10,110,210	2.000	0.000	0 (for 10); 1(for rest)	60,160,260	0.900	1.100	3
11,111,211	1.500	0.000	1	61,161,261	0.500	1.500	2
12,112,212	1.500	0.000	1	62,162,262	0.500	1.500	2
13,113,213	1.500	0.000	1	63,163,263	0.500	1.500	2
14,114,214	1.500	0.000	1	64,164,264	0.500	1.500	2
15,115,215	1.500	0.000	1	65,165,265	0.500	1.500	2
16,116,216	1.500	0.000	1	66,166,266	0.500	1.500	2
17,117,217	1.500	0.000	1	67,167,267	0.500	1.500	2
18,118,218	1.500	0.000	1	68,168,268	0.500	1.500	2
19,119,219	1.500	0.000	1	69,169,269	0.500	1.500	2
20,120,220	1.500	0.000	1	70,170,270	0.500	1.500	2
21,121,221	1.500	0.500	2	71,171,271	0.000	1.500	1
22,122,222	1.500	0.500	2	72,172,272	0.000	1.500	1
23,123,223	1.500	0.500	2	73,173,273	0.000	1.500	1
24,124,224	1.500	0.500	2	74,174,274	0.000	1.500	1
25,125,225	1.500	0.500	2	75,175,275	0.000	1.500	1
26,126,226	1.500	0.500	2	76,176,276	0.000	1.500	1
27,127,227	1.500	0.500	2	77,177,277	0.000	1.500	1
28,128,228	1.500	0.500	2	78,178,278	0.000	1.500	1
29,129,229	1.500	0.500	2	79,179,279	0.000	1.500	1
30,130,230	1.500	0.500	2	80,180,280	0.000	1.500	1
31,131,231	1.100	0.900	3	81,181,281	0.000	2.000	0 (for 81); 1(for rest)
32,132,232	1.100	0.900	3	82,182,282	0.000	2.000	0 (for 82); 1(for rest)
33,133,233	1.100	0.900	3	83,183,283	0.000	2.000	0 (for 83); 1(for rest)
34,134,234	1.100	0.900	3	84,184,284	0.000	2.000	0 (for 84); 1(for rest)
35,135,235	1.100	0.900	3	85,185,285	0.000	2.000	0 (for 85); 1(for rest)
36,136,236	1.100	0.900	3	86,186,286	0.000	2.000	0 (for 86); 1(for rest)
37,137,237	1.100	0.900	3	87,187,287	0.000	2.000	0 (for 87); 1(for rest)
38,138,238	1.100	0.900	3	88,188,288	0.000	2.000	0 (for 88); 1(for rest)
39,139,239	1.100	0.900	3	89,189,289	0.000	2.000	0 (for 89); 1(for rest)
40,140,240	1.100	0.900	3	90,190,290	0.000	2.000	0 (for 90); 1(for rest)
41,141,241	1.200	1.200	4	91,191,291	1.100	1.100	4
42,142,242	1.200	1.200	4	92,192,292	1.100	1.100	4
43,143,243	1.200	1.200	4	93,193,293	1.100	1.100	4
44,144,244	1.200	1.200	4	94,194,294	1.100	1.100	4
45,145,245	1.200	1.200	4	95,195,295	1.100	1.100	4
46,146,246	1.200	1.200	4	96,196,296	1.000	1.000	4
47,147,247	0.750	0.750	4	97,197,297	1.000	1.000	4
48,148,248	0.750	0.750	4	98,198,298	1.000	1.000	4
49,149,249	0.750	0.750	4	99,199,299	1.000	1.000	4
50,150,250	0.750	0.750	4	100,200,300	1.000	1.000	4

$^{+}$ Items were classified into one of five groups according to the latent structure: Group 1 items were loaded only on a single factor; Group 2 items were loaded primarily on one factor; Group 3 items were loaded slightly more on one factor than the other; Group 4 items were loaded equally on both factors; Group 0 items were loaded only on a single factor (like Group 1 items) and were fixed to be loaded on the corresponding factor during model estimation.

criterion (Segall, 1996). The MDPI item selection method looks for item *i* that maximizes the determinant of posterior information matrix, $|I_{i|S_{k-1}}|$, which can be expressed as

$$|I_{i|S_{k-1}}| = \Phi^{-1} + \sum_{j \in S_{k-1}} W_j + W_i, \qquad (5)$$

where Φ^{-1} is the inverse of the prior covariance matrix, S_{k-1} is the a set of administered items before *k*-th item administration, and W_i is the information matrix of item *i*. For more information about the MDPI, readers are referred to Segall (1996, 2000). For the second MAT condition, the Kullback-Leibler information (KLI) measure (Cover & Thomas, 1991; Kullback, 1959) was used as the item selection criterion. This approach was originally proposed by Chang and Ying (1996) for unidimensional CAT applications, but it's also directly applicable in multidimensional cases. For MAT, the KLI item selection criterion was defined by

$$K_i(\boldsymbol{\theta}) = \int_{\theta_0-\delta}^{\theta_0+\delta} \cdots \int_{\theta_D-\delta}^{\theta_D+\delta} K_i(\widetilde{\boldsymbol{\theta}}||\boldsymbol{\theta})\, d\widetilde{\boldsymbol{\theta}}, \qquad (6)$$

where D was the number of dimensions, and *δ* specified the range of the moving average, which was set to $3/\sqrt{k}$ in this study. $K_i(\widetilde{\boldsymbol{\theta}}||\boldsymbol{\theta})$ can be computed by

$$K_i(\widetilde{\boldsymbol{\theta}}||\boldsymbol{\theta}) = P_i(\boldsymbol{\theta}) \log\left[\frac{P_i(\boldsymbol{\theta})}{P_i(\widetilde{\boldsymbol{\theta}})}\right] + [1 - P_i(\boldsymbol{\theta})] \log\left[\frac{1-P_i(\boldsymbol{\theta})}{1-P_i(\widetilde{\boldsymbol{\theta}})}\right]. \qquad (7)$$

In practice, the integrals of Equation 6 are replaced by summations across quadrature points.

To control the item exposure rate, the fade-away (FA) method (Han, 2012) was applied after the eligible items were ordered either by the MDPI or the KLI criterion. In the FA item exposure control method, each eligible item was inversely weighted by the actual exposure rate and a target exposure rate. As a result, excessively exposed items were suppressed from item selection, whereas less used items were actively promoted for selection. A test server updated the item exposure information via a computer network.

The FA method proved effective not only in limiting the excessively exposed items but also in promoting underused items. For this study, this feature was important because it ensured similarity in the number of responses for each item across the item pool. (For more information about the FA item exposure control, readers are referred to Han, 2012). Content balancing was not implemented in the research design to avoid making it too complex. After each item administration, the interim latent trait estimate ($\hat{\theta}_{S_k}$) was computed using the maximum a posteriori (MAP; i.e., the Bayesian modal) estimation with the Newton-Raphson method.

In addition to the two MAT conditions (MDPI and KLI) described above, two other conditions also were studied as baselines. In the first of these baseline conditions, which kept all other environments the same, item selection was completely random and the missingness in the response matrix held the MCAR. For the second baseline condition, investigators generated full-response matrices (with no missing data), also keeping all other environments the same. Simulations for each condition were replicated 30 times.

Model Estimation and Evaluation

The SEM model (Figure 1) was estimated using the software package, *Mplus* 6.12 (Muthén & Muthén, 2010). For the model specification, the Group 0 items were set to load only on either F1 or F2 (Items 1 to 10 were set to load on F1, and Items 81 and 90 were set to load on F2). To avoid the indeterminacy of the latent structure and scale, the variance values for each trait were set to the true sample variances—VAR(θ_1) = 0.996 and VAR(θ_2) = 1.986—instead of fixing factor loadings (i.e., *a*-parameter) on some items to 1, for example. The mean values for the latent traits also were fixed to the true sample means—Mean(θ_1) = 0.006 and Mean(θ_2) = 0.000. The two latent traits (F1 and F2) were specified to be correlated, as shown in Figure 1, but the covariance between them was set to be estimated freely as were the item parameters (slopes and thresholds). To deal with the dichotomous responses, we used the logit link option in *Mplus*. The model also used the *marginal maximum likelihood* (i.e., "MLR" in *Mplus*) estimation with robust standard errors based on a numerical integration algorithm with a collection of iterative procedures including the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) with the Quasi-Newton and Fisher scoring optimizations.

Once the model parameters and latent scores were estimated, the parameter recovery was evaluated based on the Pearson correlation coefficients between the true parameters and estimates as well as on the bias and mean absolute error (MAE) statistics. A visual investigation on scatter plots was also conducted to evaluate the appropriateness of Pearson correlation coefficient as a

Table 2. *Recovery of Latent Trait Scores*

Estimation	Condition	Bias($\hat{\theta}_1$)	Bias($\hat{\theta}_2$)	MAE($\hat{\theta}_1$)	MAE($\hat{\theta}_2$)	Corr.($\theta_1, \hat{\theta}_1$)	Corr.($\theta_2, \hat{\theta}_2$)	Corr.($\hat{\theta}_1, \hat{\theta}_2$)
Final $\hat{\boldsymbol{\theta}}$ based on MAP	ALL	.007	.015	.123	.180	.981	.976	.706
	RAN	.016	.027	.313	.421	.912	.917	.733
	MAT(MDPI)	.015	.035	.242	.367	.946	.937	.740
	MAT(KLI)	.010	.024	.255	.364	.941	.937	.782
SEM	ALL	–.006	–.010	.121	.175	.982	.978	.718
	RAN	.000	.000	.301	.399	.913	.920	.776
	MAT(MDPI)	.001	.000	.231	.301	.951	.954	.773
	MAT(KLI)	–.012	–.020	.263	.353	.938	.941	.766

criterion. For item parameter estimates, the parameter recovery was also evaluated for each item group. To understand the more practical implications of the item parameter recovery, the multidimensional discrimination index (MDISC; Reckase & McKinley, 1991), and the multidimensional difficulty index (MDIFF; Reckase, 1985), were computed and evaluated. For the model fit comparison between the MAT condition and the "random item selection" condition (RAN), we investigated the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) statistics.

Results

MAT Administration

The final estimates on simulees' latent traits ($\hat{\boldsymbol{\theta}}$) were computed using the MAP method within the MAT administration/simulation. As reported in Table 2, the estimation biases for all four administration conditions: (a) all item administration (ALL) resulting in a full-response matrix without missing data, (b) random item selection (RAN) resulting in missingness holding MCAR, and MAT conditions (c) with MDPI and (d) with KLI resulting in missingness violating MAR, were very small (> –0.1 and < 0.1). The estimation errors based on the mean absolute error (MAE) were much smaller with ALL than with RAN, because in ALL, each simulee responded to all 300 items in the item pool, whereas each simulee in the RAN and MAT conditions responded to only 30 items. The MAEs under both MAT conditions were larger on both latent traits than the ones from ALL but smaller than the one from RAN because of the efficiency of adaptive testing. The correlation between the true parameter values and estimates also showed a similar pattern among the four ALL, RAN, and MAT conditions. The ALL condition resulted in the highest correlation between the true and the estimated $\boldsymbol{\theta}$ (0.981 and 0.976 for θ_1 and θ_2, respectively), showing the best parameter recovery performance among the studied conditions. The RAN condition, on the other hand, showed the lowest correlation coefficient (0.912 and 0.917), as one would expect when the number of test items dropped from 300 to 30. Both MAT conditions, under which each examinee was administered 30 items the same as in the RAN condition, resulted in correlation coefficients that were lower than the ones from ALL but moderately higher than RAN. Again this is indicative of improved measurement precision due to the efficiency of adaptive testing. The correlation coefficient between the estimates on the two latent traits ($\hat{\theta}_1, \hat{\theta}_2$) was very close to the true value (0.710) with ALL (0.706). With the RAN and MAT (MDPI) conditions, it was slightly overestimated (0.733 and 0.740, respectively), and with the MAT (KLI), the correlation was moderately overestimated (0.782).

Table 2 also displays the factor scores based on the SEM approach from *Mplus*. It should be noted that these factor scores were based on new item parameter estimates from the SEM analysis. The MAP estimation shown earlier did not involve estimating item parameters but rather used the item parameter data in the item pool. Unlike the final θ estimates using MAP, therefore, it is possible that the estimation errors in the factor scores from the SEM analysis potentially could have been compounded with the item parameter estimation errors.

In Table 2, the factor score results closely resembled the MAP estimation results— the score recovery was the best with ALL and the worst with RAN, and the ones with the MAT conditions were in-between. The correlation coefficients between the final estimates based on the MAP and the factor scores from the SEM were 0.98 or above across the studied conditions. The only noteworthy difference between the SEM results and the MAP results was the correlation coefficient between the estimates on the two latent traits $(\hat{\theta}_1, \hat{\theta}_2)$. When the MAP was used, the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ under RAN was 0.733, which was fairly close to the true value (0.710). When the SEM was used, the correlation under RAN changed substantially to 0.776, which could be the result of relatively larger estimation of factor scores with RAN compared with the other conditions.

In terms of item exposure control and item pool usage, each item was used 10,000 times in ALL because all items were administered to the total of 10,000 simulees. In RAN, the item exposures ranged between 994 and 1,005. In MAT (MDPI) with the FA item exposure control, the minimum exposure was 947 and the maximum exposure was 1,201, which indicated the exposure rate was effectively controlled well under the target of 0.2. In MAT (KLI) with the FA exposure control, the minimum/maximum observed exposures were 580 and 1,637, respectively, which were still well under the exposure target of 0.2. This also indicated that each test item generated at least 580 responses across the studied conditions, which were sufficient for stable SEM estimation.

SEM Estimation

There were 881 free parameters to be estimated in the SEM model, regardless of the missingness conditions. According to both the AIC and BIC index value, ALL showed the largest index values, but comparisons among RAN, MAT (MDPI), and MAT(KLI), which had similar levels of response data (i.e., same level of missingness), were the main focus. Overall, RAN resulted in a much better fit to the model than the two MAT conditions. Table 3 shows the estimated covariance between the latent traits (F1 and F2). The covariance was underestimated in all four studied conditions, and MAT (MDPI) showed the largest difference from the true sample covariance of 1.000. The other conditions (ALL, RAN, and MAT (MDPI)) came close to the true sample covariance.

Table 3. Goodness of Fit Statistics and Estimated Covariance Between Latent Traits

Condition	AIC	BIC	Estimated Covariance (F1,F2)
ALL	1875059	1881411	.976
RAN	239625	245977	.945
MAT (MDPI)	303908	310260	.988
MAT (KLI)	334756	341108	.867

The last process evaluated was the recovery of item parameters. As show in Table 4, the *d*-parameter (threshold) was reasonably well recovered under all studied conditions. The ALL and RAN conditions resulted in more than 0.99 for the correlation between the true and estimated *d*-parameter. The MAT conditions, where the MAR assumption was not held, still showed very high correlation (> .95). The scatter plot of the true and estimated *d*-parameter

Table 4. Recovery of Item Parameters by Correlation Between the True Parameter Values and Estimates

Parameter / Index	Condition	Correlation (true, est.)	Bias	MAD
d	ALL	.999	.042	0.049
	RAN	.991	.051	0.128
	MAT (MDPI)	.965	.014	0.204
	MAT (KLI)	.956	.133	0.250
$\mathbf{a}_{(1)}$	ALL	.998	-.039	0.046
	RAN	.973	-.160	0.182
	MAT (MDPI)	.926	-.033	0.164
	MAT (KLI)	.882	-.157	0.245
$\mathbf{a}_{(2)}$	ALL	.999	-.060	0.063
	RAN	.980	-.192	0.203
	MAT (MDPI)	.839	-.064	0.187
	MAT (KLI)	.793	-.228	0.310
MDISC	ALL	.978	-.082	0.082
	RAN	.762	-.284	0.288
	MAT (MDPI)	.246	-.074	0.239
	MAT (KLI)	.305	-.367	0.390
MDIFF	ALL	.999	-.017	0.031
	RAN	.993	.001	0.079
	MAT (MDPI)	.876	.019	0.131
	MAT (KLI)	.615	.001	0.196

(Figure 2), again verified the high level of *d*-parameter recovery—all item parameter estimates were extremely close to the symmetric reference line (black dashed line) regardless of the studied condition and there were no outliers that raised particular concerns.

The recovery of $\mathbf{a}_{(1)}$ and $\mathbf{a}_{(2)}$ (slope parameters) however, differed greatly across the studied conditions. With ALL, both $\mathbf{a}_{(1)}$ and $\mathbf{a}_{(2)}$ were extremely well recovered with the correlation coefficient between the true and estimated exceeding 0.99, and there was practically no estimation bias (< ± 0.1). The RAN condition also resulted in good recovery for the slope parameters ($r_{\mathbf{a},\hat{\mathbf{a}}}$ > 0.97), but the parameters overall were slightly underestimated (-.160 for $\mathbf{a}_{(1)}$ and -.192 for $\mathbf{a}_{(2)}$). As mentioned earlier, this mainly was due to the compounded problem with estimating the correlation between $(\hat{\theta}_1, \hat{\theta}_2)$ when the estimates on the latent traits were relatively less accurate compared with the other conditions (Table 2). Based on the bias and MAE statistics on $\mathbf{a}_{(1)}$ and $\mathbf{a}_{(2)}$, the estimation errors observed with RAN resulted mainly from the systematic errors. Both MAT (MDPI) and MAT (KLI) showed much lower correlation efficient values (as low as .79), indicating the MML estimation method struggled to produce stable $\mathbf{a}_{(1)}$ and $\mathbf{a}_{(2)}$ estimates when the MAR assumption was not held.

To evaluate $\mathbf{a}_{(1)}$ and $\mathbf{a}_{(2)}$ simultaneously, MDISC (often interpreted as the multidimensional discrimination index; Reckase & McKinley, 1991) also was computed for each condition. With ALL, the recovery of MDISC again was good (r = 0.978). As shown in Figure 3, ALL resulted in estimated MDISC values that were extremely close to the true values (i.e., close to the symmetric reference (dashed) line). With RAN, the correlation coefficient (*r* = .762) was lower than ALL, but, as shown in Figure 3, the deviation of the estimated MDISC looked fairly consistent despite a moderate negative bias (-0.284). With the two MAT conditions, however, the recovery of MDISC was very poor according to the correlation coefficients (0.246 and 0.305 for MAT (MDPI) and MAT (KLI), respectively). Figure 3 reveals that the estimated MDISC under both MAT conditions was very inaccurate across the level of true MDISC.

The final evaluation was of the recovery MDIFF index. The MDIFF index offers valuable information about multidimensional item difficulty (Reckase, 1985)

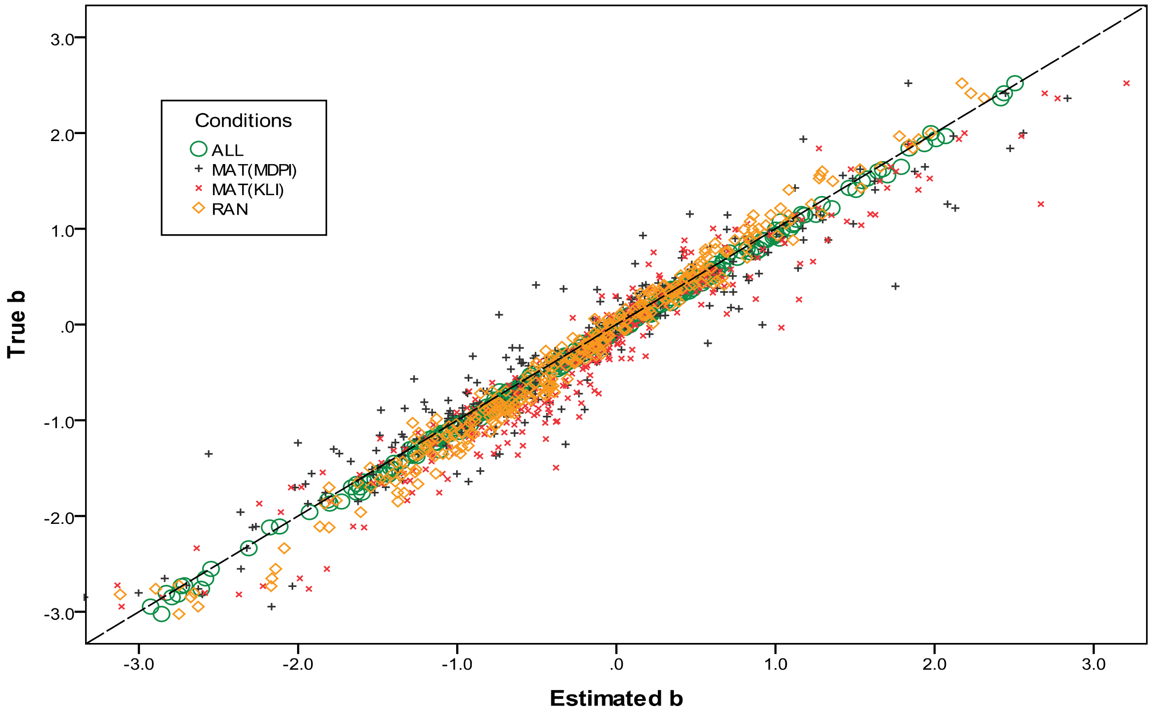

Figure 2. Recovery of b parameter values.

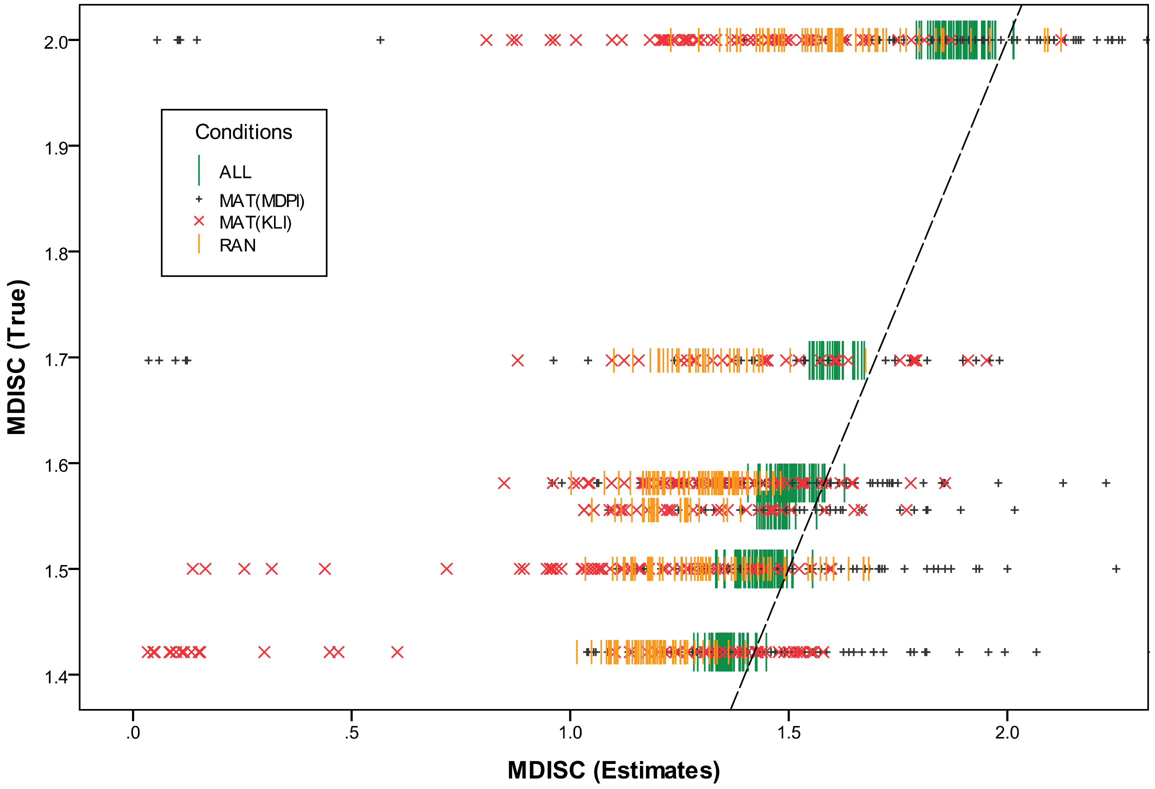

Figure 3. Recovery of MDISC values.

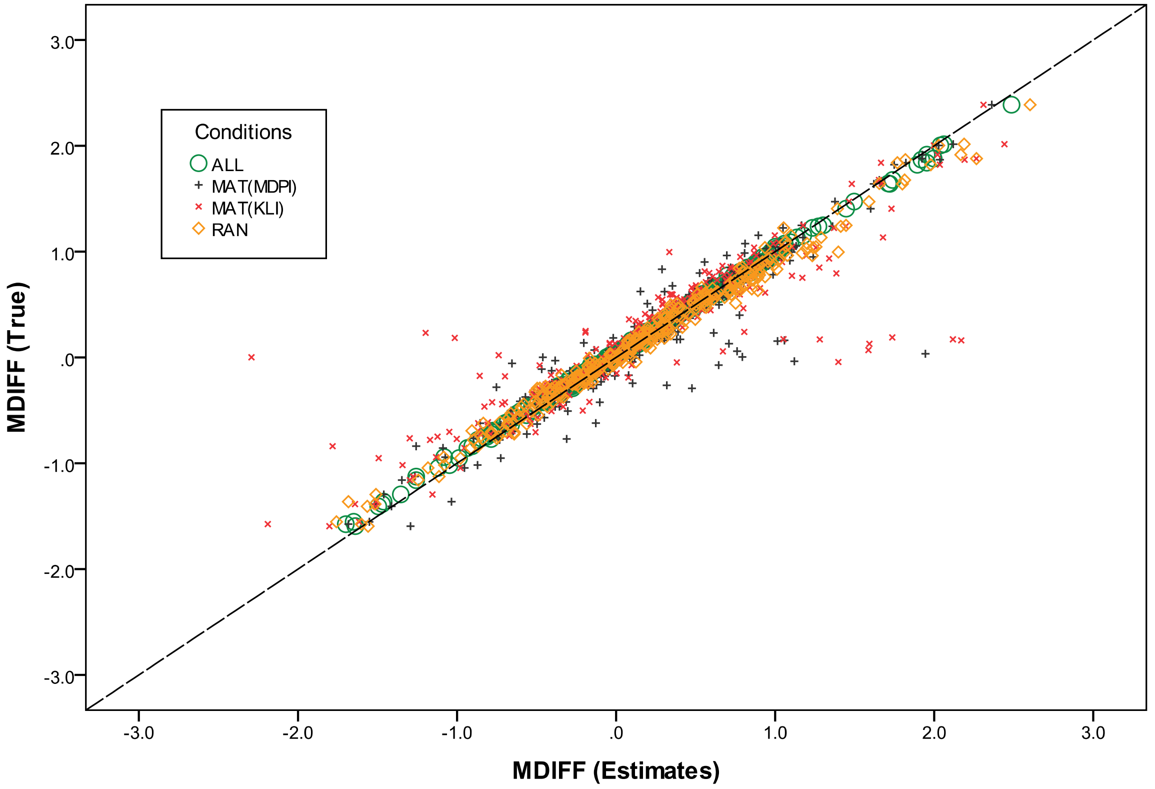

Figure 4. Recovery of MDIFF values.

and is critical as a statistic that summarizes all d-, a_((1)), and a_((2)) parameters of the MC2PL model. With ALL and RAN, the recovery of MDIFF index was nearly perfect with no meaningful estimation bias and a correlation efficient that exceeded 0.99. Under MAT (MDPI) and MAT (KLI), however, the correlation between the true and estimated MDIFF was much lower—0.876 and 0.615 for MAT (MDPI) and MAT (KLI), respectively. Interestingly, as shown in Figure 4, those items whose true MDIFF was near zero under the MAT conditions tended to have more outliers with extremely large estimation errors, and most of those outliers belonged to Group 3 or 4 and were items with similar or the same $\mathbf{a}_{(1)}$, and $\mathbf{a}_{(2)}$ parametersx

Discussion and Conclusion

Based on the results presented in Tables 2 to 4 and Figures 2 to 4, , it is apparent that the violation of the MAR assumption due to the nature of the MAT and its item selection algorithm caused the MML estimator to perform poorly, especially as seen in the comparisons of both MAT conditions to RAN. The reason for the poor performance of the MML estimation was unclear, particularly regarding the items loading similarly or equally on two factors (Groups 3 and 4) under MAT. It assuredly will require further investigation.

The findings of this study hold several important implications for MAT program developers. For (unidimensional) CAT, literature suggested that the use of adaptively administered response data for item recalibration under the MML method can be accomplished effectively with the MAR assumption (Glas, 2010; Han et al., 2011). The findings of this study suggest, however, that this may not be a valid practice for MAT. For example, some item parameter drift detection techniques involving item recalibration should not be used in MAT.

Another observation is that although the MML estimation resulted in large estimation errors under MAT, it did not necessarily lead to any meaningful systematic bias in the overall item parameters. As a result, the latent trait score estimates based on the estimated SEM model were still computed fairly accurately even under MAT conditions (Table 2). This implies that if enough items across latent traits are being measured and the majority of these items are not loaded similarly or equally on multiple factors (unlike the Group 3 or 4 items in the study), SEM-based analyses on the MAT response data may still offer fairly unbiased information about the overall test (e.g., MDIFF) and the latent structure (e.g., covariance among factors). We recommend, however, that researchers and practitioners not rely heavily on the individual item level estimates based on MAT response data because of the potential for large estimation errors. If enough data are available, multiple cross validation on parameter estimates is always recommended as it is for all other SEM-based analyses.

References

Arbuckle, J. L. (2006). Amos (Version 7.0) [Computer software]. Chicago: SPSS, Inc.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society, Series B, 42*, 293–321.

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

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: Wiley.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430–457.

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics, 13*(1), 45–52.

Glas, C. A. W. (2010). Item parameter estimation and item fit analysis. In W. J. van der Linden & C. A. W. Glass (Eds.), *Elements of adaptive testing* (pp. 269–288). New York: Springer.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Han, K. T. (2012). Efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement, 49*(3), 225–246.

Han, K. T., Guo, F., Talento-Miller, E., & Rudner, L. M. (2011, October). *Recalibrating items using adaptively administered response data.* Paper presented at the annual meeting of the International Association for Computerized and Adaptive Testing, Pacific Grove, CA.

Jöreskog, K.G. & Sörbom, D. (2006). LISREL for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Kamata, A. & Bauer, D. J. (2008). A note on the relationship between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136–153.

Kullback, S. (1959). *Information theory and statistics.* New York: Wiley.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erbaum Associates.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monograph, 15.*

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99–114.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th edition). Los Angeles: Muthén & Muthén.

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research. 1*(2), 66–82.

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

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer Science+Business Media.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension, *Applied Psychological Measurement, 14*(4), 361–373.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional space. *Psychometrika, 39*, 111–121.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* New York: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

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

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–57). Dordrecht, The Netherlands: Kluwer.

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575–588.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report 74–5). Minneapolis, MN: Psychometric Methods Program, University of Minnesota.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473–492.

Citation:

Han, Kyung T. & Fanmin Guo (2014). Impact of Violation of the Missing-at-Random Assumption on Full-Information Maximum Likelihood Method in Multidimensional Adaptive Testing. *Practical Assessment, Research & Evaluation*, 19(2). Available online: http://pareonline.net/getvn.asp?v=19&n=2.

Authors:

Kyung T. Han (corresponding)
Graduate Management Admission Council®
11921 Freedom Dr. Suite 300, Reston, VA
20190 USA
khan [at] gmac.com

Fanmin Guo
Graduate Management Admission Council®
11921 Freedom Dr. Suite 300, Reston, VA
20190 USA
fguo [at] gmac.com