The Robustness of the Unidimensional 3PL IRT Model When Applied to Two-dimensional Data in Computerized Adaptive Testing

J. Charles Zhao
Educational Testing Service

Robert F. McMorris
Robert M. Pruzek
University at Albany, State University of New York

Rusan Chen
Georgetown University

April 4, 2002

Paper presented at the 2002 Annual Meeting of the American Educational Research Association, New Orleans, LA

Acknowledgments

This study is based on the first author's doctoral dissertation. It was completed prior to his employment at Educational Testing Service.

The first author is truly indebted to his co-authors for their valuable advice and support in the process of this research endeavor. He wishes to express his gratitude to Assessment Systems Corporation for granting him a free term license to use its MicroCAT Testing System in the study. He would also like to thank Dr. Roderick P. McDonald for making the computer program NOHARM87 available for him to use in assessing the dimensionality of the generated data. Finally, thanks are due to a number of scholars for kindly answering the first author's questions regarding their published IRT-related studies when he was reviewing the literature for the research. These scholars included, in alphabetical order, Drs. Timothy N. Ansley, Frank B. Baker, Hua-Hua Chang, R. J. De Ayala, Robert A. Forsyth, Bert F. Green, Robert J. Mislevy, Martha L. Stocking, and David J. Weiss.

Author's Address

Send correspondence to J. Charles Zhao at jzhao@ets.org.

The Robustness of the Unidimensional 3PL IRT Model When Applied to Two-dimensional Data in Computerized Adaptive Testing

A revolutionary development in educational and psychological measurement in the past several decades has been the rise of item response theory (IRT). As an important application of this test theory, computerized adaptive testing (CAT) has received considerable attention from the measurement community. Built upon the unidimensional 3-parameter logistic (3PL) IRT model, CAT achieves enhanced measurement precision and reduced test length by presenting items adapted to an examinee's ability level, based on an on-going evaluation of the individual's performance during the process of testing (Wainer, 1990).

IRT models are dependent upon a set of stringent assumptions about the data to which the models will be applied. With respect to CAT, a particularly formidable assumption to meet is the one of unidimensionality, which states in essence that all the items in a test measure only one ability (Hambleton, Swaminathan, & Rogers, 1991). Although most tests nowadays are developed to measure a single ability (Weiss & Yoes, 1991), the unidimensionality assumption may prove problematic to testing practitioners simply because content and format differences in a test and various extraneous factors can easily increase the number of latent traits being assessed (Hulin, Drasgow, & Parsons, 1983; Traub, 1983). Given the practical difficulties in meeting this assumption, and a lack of appropriate CAT software to process multidimensional data, an inquiry into the robustness of the unidimensional 3PL IRT model is apparently in order. It will be especially helpful if researchers can find out how sufficiently "dominant" (Hambleton, Swaminathan, & Rogers, 1991, p. 9) a major underlying ability needs to be while the currently available CAT algorithms can still be applied without serious consequences.

Adopting a Monte Carlo approach, the measurement community began to investigate the robustness of the unidimensional 3PL IRT model to the violation of its unidimensionality assumption in the late 1970s. Using simulated multidimensional data as the true item and ability parameters, many researchers focused their attention on the effects of multidimensionality on parameter estimation by comparing the simulated parameters with their unidimensional counterparts obtained with LOGIST (Wood, Wingersky, & Lord, 1976) under systematically manipulated conditions.

Reckase (1979) generated data to fit a linear factor-analytic model for the simulation part of his study. It was reported that when there were several equally potent factors, the ability estimates were highly correlated with the factor scores for just one of such factors; when there was a dominant first factor, the ability estimates were highly correlated with the factor scores for that first factor.

Drasgow and Parsons (1983) employed a hierarchical factor-analytic model to generate multidimensional data in their study. They found that, whether guessing was involved in item responses or not, with first-order common factors correlated from .46 to .90, estimates of the item and ability parameters were closely related to the parameters associated with the second-order general factor. When the potency of the second-order general factor decreased, these estimates became more closely related to the parameters associated with the most potent first-order common factor instead.

Ansley and Forsyth (1985) adopted as their simulation model the two-dimensional version of Sympson's (1978) noncompensatory multidimensional IRT (MIRT) model with the c_i value fixed at .2. They reported that, regardless of the sample size and the test length, the estimates of ability parameters ($\hat{\theta}$) seemed best considered the average of the true θ_1 and θ_2 values. However, a substantial disparity always existed between the magnitudes of the statistics derived from the two-dimensional data and those derived from the unidimensional data.

Doody (1985) studied the IRT model robustness issue in a vertical equating context, using the two-dimensional version of Doody-Bogan and Yen's (1983) compensatory MIRT model. With a cross-validation component incorporated in her design, the researcher found that the use of the unidimensional 3PL IRT model in parameter estimation was as good for multidimensional data as it was for unidimensional data in most of the test conditions simulated in her study. Yet, she warned in test equating terms that the poorest item parameter estimates would occur when one test was unidimensional and one test was multidimensional.

Way, Ansley, and Forsyth's (1988) study involved the use of the two-dimensional version of both Sympson's (1978) and Doody-Bogan and Yen's (1983) MIRT models with c_i values fixed at .2. These researchers verified Ansley and Forsyth's (1985) findings regarding the ability estimates, and further noticed that the relationship between $\hat{\theta}$ and the average of θ_1 and θ_2 varied differently with the relationship between θ_1 and θ_2 across MIRT models.

Ackerman (1989) employed in his study the two-dimensional versions of McKinley and Reckase's (1982) compensatory and Sympson's (1978) noncompensatory MIRT models with c_i values fixed at 0. He found that the $\hat{\theta}$ values were about equally correlated with the θ_1 and θ_2 values across all $r_{\theta_1\theta_2}$ levels for both models. Ackerman also examined the differences between LOGIST (Wingersky et al., 1982) and BILOG (Mislevy & Bock, 1982) in his study.

The measurement community did not begin to assess the robustness of the unidimensional 3PL IRT model in various CAT settings until the early 1980s. In this new research endeavor, simulated examinees (simulees) with known multidimensional abilities were adaptively tested based on a unidimensional 3PL IRT model, and their unidimensionally derived ability estimates were then compared with their true multidimensional ability levels.

. Weiss and Suhadolnik (1982) co-authored an early study in which multidimensional data were generated based on the factor structures of the ASVAB General Science subtest, and the ability estimates were obtained with Birnbaum's (1968) maximum likelihood estimation method. They found that, with the increase of multidimensionality in the generated item responses, the $\hat{\theta}$ values departed further from the true first factor θ values across all the levels evaluated. However, the effects of multidimensionality could be overcome by increasing the test length in many cases.

Folk and Green (1989) included a partial replication component in their study and used the two-dimensional version of Hattie's (1981) compensatory MIRT model to generate data. The item parameters were calibrated with a modified maximum likelihood estimation method, and the ability estimates were obtained through a combination of Owen's (1975) Bayesian sequential estimation method and Samejima's (1969) Bayesian modal estimation method. Folk and Green

reported that, with the item set containing mutually exclusive subsets of items that each measure one ability only, the $\hat{\theta}$ values tended to be close to either θ_1 or θ_2 value as the $r_{\theta_1\theta_2}$ level decreased. With the item set measuring differing amounts of each ability across items, the $\hat{\theta}$ values were equally related to the θ_1 and θ_2 values regardless of the $r_{\theta_1\theta_2}$ levels.

De Ayala (1992) generated data for his study using the two-dimensional version of Doody-Bogan and Yen's (1983) compensatory MIRT model with the c_i value set to .20. Based on a unidimensional item pool, each simulated CAT session in his study estimated the simulee's ability level using Owen's (1975) Bayesian sequential estimation method. De Ayala found that the increased association of difficulty parameters improved the accuracy of ability estimation slightly. The correlation between $\hat{\theta}$ and the average of θ_1 and θ_2 values increased as the associations between interdimensional abilities became stronger. In all situations, $r_{\hat{\theta}\bar{\theta}}$ became greater when the a_1 and a_2 values in the item pool had been sorted in opposite directions.

A review of the literature on the robustness of the unidimensional 3PL IRT model shows that all researchers noticed the effects of multidimensionality on ability estimation. However, they differed as to the characteristics of the unidimensional estimates of the multidimensional abilities. The purpose of this Monte Carlo study was to investigate the robustness of the unidimensional 3PL IRT model to the violation of its unidimensionality assumption in CAT applications. Through a replication and extension of previous research in this area, the current study was designed to address the following research questions:

1. Do different correlations between the two item difficulty dimensions ($\rho_{b_1b_2}$s) affect CAT results?

2. Do different correlations between the two ability dimensions ($\rho_{\theta_1\theta_2}$s) affect CAT results?

3. Do different ability estimation methods (METHOD) affect CAT results?

4. Do the above three factors interact to affect CAT results?

5. How do CAT results differ when they are based on unidimensional and two-dimensional item response data respectively?

Method

Computer Programs

Several computer programs were employed to facilitate this Monte Carlo study. They included (a) IRMG (Item Response Matrix Generator) for producing all the required data sets, (b) NOHARM87 (a more recent version of NOHARM, Fraser & McDonald, 1988) for verifying the dimensionality of the generated item response data sets, (c) ASCAL (Vale & Gialluca, 1985) for estimating item parameters in every item bank, and (d) SimuCAT for CAT simulation. ASCAL was a component of the MicroCAT Testing System Version 3.0 (1993) (henceforth

referred to as MicroCAT), and IRMG and SimuCAT were developed by the first author of this study in SAS Interactive Matrix Language (SAS/IML) (SAS Institute, 1990). SimuCAT modeled on MicroCAT in testing algorithm and ability estimation methods.

Simulation Models

A modified version of Birnbaum's (1968) 3PL IRT model was selected for item and ability estimations in this study. This version can be expressed as

$$P_i(\theta) = c + \frac{1-c}{1+\exp[-1.7a_i(\theta - b_i)]} \qquad (i = 1, 2, \ldots, n), \tag{1}$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with ability answers item i correctly,

a_i is the discrimination parameter for item i,

b_i is the difficulty parameter for item i,

c is a constant across items as the pseudo-chance-level parameter, and

n is the number of items in the test.

To generate multidimensional item response matrices, the two-dimensional versions of two different MIRT models were used. These models included Doody-Bogan and Yen's (1983) compensatory model:

$$P_{ij}(\theta_{jk}) = c_i + \frac{1-c_i}{1+\exp[-1.7(\sum_{k=1}^{m} a_{ik}(\theta_{jk} - b_{ik})]} \qquad (i,k = 1, 2,\ldots,n;\ j = 1,2,\ldots, N), \tag{2}$$

where $P_{ij}(\theta_{jk})$ is the probability of a correct response to item i by examinee j whose location in an m-dimensional latent space is described by the ability vector θ_{jk},

a_{ik} is the discrimination parameter for item i on dimension k,

b_{ik} is the difficulty parameter for item i on dimension k,

N is the number of examinees who respond to item i, and

all other parameters are as defined previously,

and Sympson's (1978) noncompensatory model:

$$P_{ij}(\theta_{jk}) = c_i + \frac{1 - c_i}{\prod_{k=1}^{m} \{1 + \exp[-1.7 a_{ik}(\theta_{jk} - b_{ik})]\}} \qquad (i, k = 1, 2, ..., n;\ j = 1, 2, ..., N), \tag{3}$$

where all parameters have already been defined.

Basic Research Design

Item Response Data

Item parameters. Five two-dimensional item pools with 246 items each were created for this study. The a_1 values were randomly sampled from a uniform distribution in the range [0.5, 2.0]. To simulate a standardized achievement test that primarily tapped the ability on the first dimension, the a_2 values were randomly sampled from a uniform distribution in the range [0.25, 1.0] such that the a_1 values would be twice as large as the a_2 values on average. These two *a* vectors were then used in the entire five item pools.

The b_1 values of every item pool were evenly distributed at 41 points in the range [-2.0, 2.0] with every six items sharing the same point, so that for the first six items $b_1 = -2.0$, for the second six items $b_1 = -1.9$, and so on. To make the b_1 and b_2 values correlated differently across the five item pools ($\rho_{b_1 b_2}$ = .1, .45, .63, .77, and .89), a 246×1 vector *u* was created for every item pool to fit a uniform distribution in the range [0, 1]. The initial values of the b_2 vector were then computed by using Hoffman's (1959) equation:

$$b_2 = b_1 + \frac{\sqrt{1 - \rho_{b_1 b_2}^2}\, \sigma_{b_1} u}{\rho_{b_1 b_2} \sigma_u}, \tag{4}$$

where σ_{b_1} is the standard deviation of the b_1 vector,

σ_u is the standard deviation of the *u* vector, and

all other parameters are as defined previously.

To emphasize the primary status of the abilities on the first dimension and to keep the b_2 values within the desired range, the initial b_2 values were linearly transformed, so that their mean was 1 less than the mean of their corresponding b_1 values, and their standard deviation was at 0.4. In every item pool, the c_i values were fixed at .2.

Some characteristics of the item parameters need explanation. The pool size was set at 246 because the distribution of the b_1 values required it to be a multiple of 6, and the maximum

processing capacity of the item calibration program ASCAL (Vale & Gialluca, 1985) was 250 items. The distribution of the a_1 and b_1 values and the use of Hoffman's (1959) method to generate the b_2 values were adapted from De Ayala (1992). The range of the a_1 values came from Ansley and Forsyth (1985), Doody (1985), and Way et al. (1988). The range of the b_1 values was due to Ansley and Forsyth, Folk and Green (1989), and Way et al. The *c* value was obtained from Drasgow and Parsons (1983), Ansley and Forsyth, and De Ayala. The correlations between the b_1 and b_2 values were selected to cover a series of evenly spread linear relationships between the *b* vectors (i.e., $.1^2 = .01 \approx .00$, $.45^2 = .2$, $.63^2 = .4$, $.77^2 = .6$, and $.89^2 = .8$), assuming that squared correlation is a generally accepted measure of linear predictability in terms of the variance accounted for. Finally, the desired correlation of .1 was chosen to approximate the correlation of 0, which is not defined in Hoffman's method.

S*imulees*. Five two-dimensional 2,000-simulee samples (subsequently referred to as Simulee Set A) were generated for item calibration. As specified by Ansley and Forsyth (1985), Doody (1985), Way et al. (1988), Folk and Green (1989), and Ackerman (1989), the vectors for each sample were produced to approximate a bivariate normal distribution with zero mean and unit variance for each dimension. Different correlations between the vectors ($\rho_{\theta_1\theta_2}$ = 0, .45, .63, .77, and .89) were obtained across the five samples by using the bivariate version of Moonan's (1957) formula

$$Z = AX, \tag{5}$$

where *A* is a lower triangular 2×2 matrix that results from a Cholesky factorization of the matrix *M,* which depicts the desired correlation between the vectors,

X is the initial $2 \times 2{,}000$ matrix of independent pseudo-random normal deviates with expected zero mean and unit variance for rows, and

Z is a $2 \times 2{,}000$ matrix of linearly transformed pseudo-random normal deviates with expected zero mean and unit variance for rows, and the desired correlation between the vectors.

Like $\rho_{b_1b_2}$, the $\rho_{\theta_1\theta_2}$s chosen here were intended to represent a number of evenly distributed linear relationships between the vectors that could be used to facilitate linear prediction.

Simulee Set B, which included five two-dimensional 1,000-simulee samples, was drawn in the same manner with a different seed value for CAT simulations.

Item response matrices. Combinations of $\rho_{b_1b_2}$ (.1, .45, .63, .77, and .89), $\rho_{\theta_1\theta_2}$ (0, .45, .63, .77, and .89), and MIRT model (compensatory and noncompensatory) produced a total of fifty $2{,}000 \times 246$ matrices of probabilities of a correct response with elements p_{ij}. To introduce a random error component into every matrix, a random number *r* was drawn from a uniform distribution in the range [0, 1] for every p_{ij} value. By applying the rule

$$x_{ij} = \begin{cases} 1 \text{ (a correct answer) if } p_{ij} \geq r \\ 0 \text{ (an incorrect answer) if } p_{ij} < r \end{cases}, \quad (6)$$

fifty (0, 1) matrices with elements x_{ij} were created. Based on Simulee Set A, these dichotomous item response data sets were used for item calibration.

Based on Simulee Set B, 50 additional (0, 1) matrices with elements x_{ij} were generated in the same manner. Involving 1,000 simulees each, these data sets were used as simulee responses when adaptive testing was simulated.

Item Banks

Item calibrations using the 50 Simulee Set A-based item response matrices were completed according to Vale and Gialluca's (1985) pseudo-Bayesian method:

$$L^* = \prod_{j=1}^{N} \prod_{i=1}^{n} P_{ij}^{v_{ij}} Q_{ij}^{1-v_{ij}} \Phi(\theta_j) f(a_i, 3.0, 3.0, 0.3, 2.6) f[c_i 5.0, 5.0, -0.05, (0.05 + 2/K)], \quad (7)$$

where L^* is the pseudo-likelihood function for omitted item responses,

Q_{ij} is the probability of an incorrect response to item i by examinee j,

$\Phi(\cdot)$ is the standard normal cumulative density function,

$$v_{ij} = \begin{cases} 1 \text{ if examinee } j\text{'s response to item } i \text{ is correct,} \\ 0 \text{ if examinee } j\text{'s response to item } i \text{ is incorrect,} \\ 1/K \text{ otherwise, where K is the number of alternatives, and} \end{cases}$$

$f(x, r, s, l, u)$ is the Bayesian prior for parameter x (a_i or c_i), with upper and lower bounds u and l, and beta function parameters r and s. (Bounded by ± 3.0, b_i has no prior distribution.), and

all other parameters are as defined previously.

This led to the creation of 50 item banks with unidimensional item parameter estimates: 25 compensatory, and 25 noncompensatory. To ensure a high quality of the test items for CAT simulations, these item banks were subsequently cleaned by deleting items whose parameter estimates failed to converge to their true values. As recommended by Assessment Systems Corporation' (1994), items were also removed if their χ^2 lack-of-fit statistics were about four times as large as those of the other items in the same bank.

CAT Simulations

The maximum information item selection strategy was adopted in all CAT simulations. To expedite the process of simulated test administration, an information lookup table was created

for each of the 50 item banks at each of the 81 equally distributed ability levels in the range [-4.0, 4.0]. These tables were based on Birnbaum's (1968) item information function:

$$I_i(\theta) = \frac{2.89a_i^2(1-c_i)}{\{c_i + \exp[1.7a_i(\theta - b_i)]\}\{1+\exp[-1.7a_i(\theta - b_i)]\}^2}, \tag{8}$$

where $I_i(\theta)$ is the information provided by item *i* at θ, and

all other parameters are as defined previously.

All the simulees from Simulee Set B were adaptively tested against their corresponding item banks. Four ability estimation methods were used. They included Birnbaum's (1968) maximum likelihood estimation:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{\left[\frac{\partial \ln L}{\partial \theta}\right]_t}{\left[\frac{\theta^2 \ln L}{\partial \theta^2}\right]_t} = \hat{\theta}_t - \frac{\left[\sum_{i=1}^{n} \frac{1.7a_i(u_i - P_i)(P_i - c_i)}{P_i(1-c_i)}\right]_t}{\left[-\sum_{i=1}^{n} \frac{2.89a^2 Q_i (P_i - c_i)^2}{(1-c_i)_{\mathrm{i}}^2 P_{\mathrm{i}}}\right]_t}, \tag{9}$$

where t is the iteration index,

L is the likelihood function $L = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}$,

u_i is the response of a randomly selected examinee to item *i*, coded 1 for a correct response and coded 0 for an incorrect response, and

all other parameters are as defined previously,

Owen's (1975) Bayesian sequential estimation:

$$\mathrm{E}(\theta|1) = \mu + \frac{(1-c_i)\sigma^2 \phi(D)}{\sqrt{\frac{1}{a_i^2} + \sigma^2 A}}, \tag{10.1}$$

$$\mathrm{E}(\theta|0) = \mu - \frac{\sigma^2 \phi(D)}{\sqrt{\frac{1}{a_i^2} + \sigma^2 \Phi(D)}}, \tag{10.2}$$

where $\mathrm{E}(\theta \mid 1) / \mathrm{E}(\theta \mid 0)$ is the posterior mean given a correct response (1) or an incorrect response (0),

μ and σ are the mean and standard deviation of the prior distribution,

$$A = c_i + (1 - c_i)\Phi(-D),$$

$$D = \frac{b_i - \mu}{\sqrt{\frac{1}{a_i^{\,2}} + \sigma^{\,2}}}, \text{ and}$$

$\phi(\cdot)$ is the standard normal probability density function,

Samejima's (1969) Bayesian modal estimation:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{\left[\frac{\partial \ln L}{\partial \theta}\right]_t}{\left[\frac{\partial^{\,2} \ln L}{\partial \theta^{\,2}}\right]_t} = \hat{\theta}_t - \frac{\left[\sum_{i=1}^{n} \frac{1.7 a_i (u_i - P_i)(P_i - c_i)}{P_i (1 - c_i)} - \left(\frac{\hat{\theta}_t - \mu}{\sigma^{\,2}}\right)\right]_t}{\left[-\sum_{i=1}^{n} \frac{2.89 a^2 Q_i (P_i - c_i)^2}{(1 - c_i)_{\mathrm{i}}^2 P_{\mathrm{i}}} - \frac{1}{\sigma^{\,2}}\right]_t}, \tag{11}$$

where $\sigma^{\,2}$ is the variance of the prior θ distribution, and

all other parameters are as defined previously,

and Bock and Aitken's (1981) Bayes EAP estimation:

$$\mathrm{EAP}(\theta) = \mathrm{E}(\theta | \mathbf{u}) = \frac{\sum_{k=1}^{q} L(\mathbf{u} | X_k) A(X_k) X_k}{\sum_{k=1}^{q} L(\mathbf{u} | X_k) A(X_k)}, \tag{12}$$

where $L(\mathbf{u} \mid X_k)$ is the log-likelihood of the observed response pattern $\mathbf{u}$ $(u_1, u_2, \ldots, u_n)$, given the *k*th Gauss-Hermite node X_k, and

$A(X_k)$ is the weight of X_k.

A simulee's unidimensional ability estimate $\hat{\theta}$ was fixed at 0 at the beginning of each of the 200,000 simulated CAT sessions (50 item banks $\times 4$ ability estimation methods $\times 1{,}000$ simulees). After an item with the highest level of information at $\hat{\theta}$ was selected from the information lookup table and administered, the simulee's new $\hat{\theta}$ was calculated based on his/her item response retrieved from the corresponding item response matrix. This estimation process would continue until either a maximum of 30 items was administered, or a standard error of no more than 0.05 was obtained after the administration of a minimum of 20 items.

During the CAT simulations, most of the ability estimation methods functioned under certain specifications. For maximum likelihood and Bayesian modal estimation, which entail iterations of the Newton-Raphson procedure, the absolute increment to a $\hat{\theta}$ was restricted to 0.5. As suggested by Baker (1992), the iterations were terminated when either an iterative cycle of 20

was completed or the absolute increment to a $\hat{\theta}$ became less than 0.001. Regarding Bayesian sequential and Bayesian modal estimation, a normal prior distribution of was chosen that has zero mean and unit variance.

Figure 1 presents the basic structure of a two-dimensional data-based CAT simulation as a system flowchart. It should be pointed out that despite the use of unidimensional item parameter estimates for item selection and ability estimation in the CAT simulations, simulee responses were always created from a two-dimensional model.

Unidimensional Data-based Simulations

Unidimensional data-based CAT simulations were carried out to establish a frame of reference for detecting errors associated with the ability estimation methods when the results of the two-dimensional data-based CAT simulations were analyzed.

Five unidimensional item pools with 246 items each were generated for the planned CAT simulations based on Birnbaum's (1968) 3PL IRT model. For every item pool, the *a*, *b*, and *c* values were generated in the same manner as the a_1, b_1, and *c* values in the two-dimensional item pools, and the two corresponding simulee samples for item calibration $(N = 2{,}000)$ and CAT simulation $(N = 1{,}000)$ were each drawn from a normal distribution with zero mean and unit variance. All the simulation procedures remained basically the same as those used in the two-dimensional data-based simulations except that the MIRT models were replaced by the unidimensional IRT model.

Two different seed values were employed for each of the five unidimensional data-based CAT simulations. One was used to generate the *a* and values for item calibration, and the other the values for CAT simulation.

To obtain more stable results from this Monte Carlo study, the basic research design described above was replicated five times.

Data Analysis

Adequacy of Certain Computer Programs

The data analysis began with a validation of the SAS/IML programs IRMG and SimuCAT. The validity of the generated *a*, *b*, and values was evaluated by comparing the differences between the expected and the observed values of their means, standard deviations, ranges, and correlations where applicable. The dimensionality of the item response matrices was verified by computing the root mean square of residuals (RMSR) for a number of representative matrices with NOHARM87. The accuracy of the CAT simulations was determined by finding the differences between SimuCAT and MicroCAT in terms of the ability estimates and their variances based on a sample item response matrix.

Robustness of Unidimensional 3PL IRT Model in CAT

A number of statistical indices were calculated at the simulee sample level to evaluate the CAT simulation results under various test conditions.

The fidelity indices included Pearson product-moment and Spearman rho correlations between $\hat{\theta}$ and θ_1 ($r_{\hat{\theta}\theta_1}$ and $rs_{\hat{\theta}\theta_1}$), between $\hat{\theta}$ and θ_2 ($r_{\hat{\theta}\theta_2}$ and $rs_{\hat{\theta}\theta_2}$), and between $\hat{\theta}$ and the average of θ_1 and θ_2 ($r_{\hat{\theta}\bar{\theta}}$ and $rs_{\hat{\theta}\bar{\theta}}$).

The bias index can be expressed as

$$Bias_k = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_{jk})}{N} \qquad (k = 1, 2, 3), \tag{13}$$

where θ_{jk} becomes the average of θ_1 and θ_2 for examinee *j* when k = 3, and

all other parameters are as defined previously.

The error indices included root mean square error (RMSE) and average standard error of estimation (ASE). They were computed as follows:

$$RMSE_k = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_{jk})^2}{N}} \qquad (k = 1, 2, 3), \tag{14}$$

where all parameters are as defined previously.

$$ASE = \frac{\sum_{j=1}^{N} SE(\hat{\theta})_j}{N}, \tag{15}$$

where $SE(\hat{\theta})_j$ is the standard error of estimation associated with $\hat{\theta}_j$ when a testing session was terminated, and *N* is as defined previously. Depending on the ability estimation method, $SE(\hat{\theta})_j$ can be obtained by

$$SE(\hat{\theta})_j = \sqrt{\frac{1}{-\mathrm{E}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}} = \sqrt{\frac{1}{-\sum_{i=1}^{n}\frac{2.89a^2Q_i(P_i - c_i)(u_i c_i - P_i^2)}{(1-c_i)_i^2 P_i^2}}} \tag{16}$$

(Maximum likelihood estimation),

$$SE(\hat{\theta}\mid 1)_j = \sqrt{\sigma^2\left\{1-\frac{(1-c_i)\phi(D)\left[\frac{(1-c_i)\phi(D)}{A}-D\right]}{\left(1+\frac{1}{a_i^2\sigma^2}\right)A}\right\}}, \tag{17.1}$$

$$SE(\hat{\theta}\mid 0)_j = \sqrt{\sigma^2\left\{1-\frac{\phi(D)\left[\frac{\phi(D)}{\Phi(D)}+D\right]}{\left(1+\frac{1}{a_i^2\sigma^2}\right)\Phi(D)}\right\}} \tag{17.2}$$

(Bayesian sequential estimation),

$$SE(\hat{\theta})_j = \sqrt{\frac{1}{-\mathrm{E}\left[\frac{\partial^2 \ln L}{\partial\theta^2}\right]+\frac{1}{\sigma^2}}} = \sqrt{\frac{1}{-\sum_{i=1}^{n}\frac{2.89a^2Q_i(P_i-c_i)(u_ic_i-P_i^2)}{(1-c_i)_i^2P_i^2}+\frac{1}{\sigma^2}}} \tag{18}$$

(Bayesian modal estimation), or

$$\mathrm{SE}(\hat{\theta})_j = \sqrt{\frac{\sum_{k=1}^{q}(X_k-\hat{\theta})^2L(\mathbf{u}\mid X_k)A(X_k)}{\sum_{k=1}^{q}L(\mathbf{u}\mid X_k)A(X_k)}} \tag{19}$$

(Bayes EAP estimation),

where all parameters have already been defined.

The test efficiency index was defined as the average number of items administered in a testing session:

$$ANI = \frac{\sum_{j=1}^{N}NI_j}{N}, \tag{20}$$

where NI_j is the number of items used when testing session *j* is terminated, and

N is as defined previously.

Magnitudes of Factorial Effects

This Monte Carlo study was based on three simulation factors: $\rho_{b_1 b_2}$, $\rho_{\theta_1 \theta_2}$, and METHOD. The magnitudes of the effects of these simulation factors were assessed using analysis of variance (ANOVA) with planned comparisons. $RMSE_3$ was selected as the dependent variable for the analysis. This was not only because RMSE, in general, was sensitive to changes in treatment conditions in an experimental design, but also because $RMSE_3$, as will be shown later, could capture the characteristics of the unidimensionally estimated two-dimensional θs better than $RMSE_1$ and $RMSE_2$. The logarithm of $RMSE_3$ ($LRMSE_3$) was taken to improve the model-data fit and help normalize the residuals.

In this three-way ANOVA, the linear and quadratic trends related to $\rho_{b_1 b_2}$ and $\rho_{\theta_1 \theta_2}$, the main effect of METHOD, as well as all the relevant interactions were each explored by creating a system of contrast coding. These coding systems included two vectors of orthogonal polynomial coefficients for both $\rho_{b_1 b_2}$ and $\rho_{\theta_1 \theta_2}$

[2 1 0 -1 -2] (Linear)

[2 -1 -2 -1 2] (Quadratic),

three vectors of orthogonal polynomial coefficients (Helmert contrasts) for METHOD

[3 -1 -1 -1] (Maximum likelihood vs. all Bayesian estimation methods)

[0 2 -1 -1] (Bayesian sequential vs. other Bayesian estimation methods)

[0 0 1 -1] (Bayesian modal vs. Bayes EAP estimation methods),

and 28 vectors of derived orthogonal coefficients corresponding to the interactions, which were constructed by cross-multiplying the appropriate $\rho_{b_1 b_2}$, $\rho_{\theta_1 \theta_2}$, and METHOD vectors. A measure of the magnitude of the effect of a simulation factor was obtained by computing the squared Pearson product-moment correlation coefficient between the dependent variable and each relevant vector of orthogonal coefficients. In all cases, ANOVA with planned comparisons was used as a descriptive method to quantify effects of the simulation factors. Because a "significant" trend or main effect could well be the artifact of the number of replications arbitrarily chosen for such a Monte Carlo study, an inferential use of the ANOVA technique was not warranted in this situation.

Results

Adequacy of Certain Computer Programs

Quality of Generated Data

Validity in terms of item and ability characteristics. The summary statistics for the generated item and ability parameters in Tables 1 and 2 were averaged across the basic research design and the five subsequent replications. These statistics show that the generated data fulfilled all the criteria previously specified.

Validity in Terms of Item Response Matrix Dimensionality. To assess the dimensionality of the item response matrices based on the generated two-dimensional item and ability

parameters, six representative matrices were selected for nonlinear factor analysis. Because of the processing capacity limitations of NOHARM87, only a random sample of 123 items were retained in each selected matrix when its dimensionality was examined. Table 3 contains the RMSR values generated by HOHARM87 after conducting exploratory analyses of the relevant matrices with the convergence criterion set to the default value of 0.000001. For matrices r1nc1000, r2co4545, and r4co7777, the percentages of RMSR value reduction (i.e., the percentages of improvement in model-data fit) are notably greater between the 1-factor and the 2-factor solutions (19%, 24%, and 26%) than between the 2-factor and the 3-factor solutions (3%, 5%, and 2%). This can be regarded as a sign that a 2-factor solution can better capture the dimensionalities of the item response matrices in question. For the other three matrices cited in Table 3, however, their dimensionalities are not so clear. Although the relatively small percentage reduction (8%) between the 1-factor and the 2-factor solutions with the matrix r5nc8989 might result from a high correlation between the θ_1 and θ_2 vectors (.89), the percentage reductions between the 1-factor and the 2-factor solutions with the matrix r3nc6363 (6%), and between the 2-factor and the 3-factor solutions with the matrix baco1000 (17%) were apparently out of their respective expected ranges in such a context.

Quality of CAT Simulations

The adequacy of SimuCAT is established if SimuCAT can be found virtually identical to MicroCAT in terms of the ability estimates and the related variances the two of them produced. A random sample of 50 items and 100 simulees associated with the first unidimensional item response matrix for the basic research design was selected for a comparative study of SimuCAT and MicroCAT. All the CAT simulation procedures specified for this research were followed in the study except the stopping rule. To accelerate the simulation speed, the maximum and the minimum numbers of items to be administered per CAT session, as well as the maximum variance allowed when a CAT session was terminated were reduced to 25, 15, and 0.1 respectively. For the MicroCAT run, a simulee's item response was manually entered at every interactive CAT session. Table 4 displays the differences between the results of the SimuCAT and MicroCAT runs. The SAS/IML program and the commercial software package appeared to be in perfect agreement on Bayesian sequential and Bayes EAP estimation results. Yet, they differed, to some extent, in estimation results involving the maximum likelihood and the Bayesian modal estimation methods. These differences could be traced back to the fact that SimuCAT could not utilize the copyright-protected ad hoc decision rules and special estimation techniques incorporated in MicroCAT for the implementation of these estimation methods. Given the relatively small size of these differences, SimuCAT can be deemed adequate for CAT simulation purposes.

Robustness of Unidimensional 3PL IRT Model in CAT

The results of the CAT simulations across MIRT and IRT models presented in this subsection were based on item banks that were cleaned according to the statistical criteria previously specified. On average, 2.7 items were eliminated from each compensatory item pool, 11.8 items from each noncompensatory item pool, and 7.5 items from each unidimensional item pool. Regarding the CAT simulations, the item usage rates reached approximately 67%, 65%,

and 73% per item bank for compensatory, noncompensatory, and unidimensional data respectively.

In the data tables presented in this subsection, an evaluative index associated with each level of a simulation factor was averaged across all combinations of the individual levels of the other two simulation factors in the basic research design and the five subsequent replications. Possible interactions between every two simulation factors were investigated by examining the evaluative indices averaged across all levels of a third simulation factor.

$\rho_{b_1 b_2}$ *Effect*

The evaluative indices in Table 5 reflect the CAT simulation results associated with the various levels of $\rho_{b_1 b_2}$. Across MIRT models, the $\hat{\theta}$ values were highly correlated with the θ_1 and $\bar{\theta}$ values. The observed Pearson product-moment correlation coefficients $r_{\hat{\theta}\theta_1}$, $r_{\hat{\theta}\theta_2}$, and $r_{\hat{\theta}\bar{\theta}}$ remained virtually unchanged over the decreasing $\rho_{b_1 b_2}$. Although $r_{\hat{\theta}\theta_1}$ was apparently larger than $r_{\hat{\theta}\theta_2}$, $r_{\hat{\theta}\bar{\theta}}$ always exceeded $r_{\hat{\theta}\theta_1}$ to some extent. As special measures of bivariate rank order association, the observed Spearman rho correlation coefficients $rs_{\hat{\theta}\theta_1}$, $rs_{\hat{\theta}\theta_2}$, and $rs_{\hat{\theta}\bar{\theta}}$ turned out to be very similar to their Pearson product-moment counterparts. An inspection of the values of $Bias_1$ through $Bias_3$ shows that the CAT simulation results basically represented a negligible overestimate of the true ability on every dimension. This type of bias changed little with the decrease of $\rho_{b_1 b_2}$. Of the three bias measures, $Bias_1$ was the smallest, and $Bias_2$ the largest. With a few exceptions with the noncompensatory data only, the *RMSE*s and *ASE* experienced a systematic yet inconsequential increase with the decrease of $\rho_{b_1 b_2}$, and the *RMSE*s moved upward somewhat noticeably when $\rho_{b_1 b_2}$ dropped from .45 to .1. $RMSE_1$ and $RMSE_3$ were each about half as large as $RMSE_2$, and $RMSE_3$ appeared to be consistently smaller than $RMSE_1$. The increase in *ANI* was small as b_1 and b_2 became increasingly uncorrelated. Except for the correlation coefficients, evaluative indices associated with $\rho_{b_1 b_2}$ appeared larger with the noncompensatory data than with the compensatory data.

$\rho_{\theta_1 \theta_2}$ *Effect*

Table 6 displays the evaluative indices of the CAT simulation results across all levels of $\rho_{\theta_1 \theta_2}$. For both compensatory and noncompensatory MIRT models, there was a close relationship between the $\hat{\theta}$ and θ_1 values and between the $\hat{\theta}$ and $\bar{\theta}$ values. The observed Pearson product-moment correlation coefficients diminished as $\rho_{\theta_1 \theta_2}$ decreased. This change was more pronounced when $\rho_{\theta_1 \theta_2}$ fell from .45 to 0. Except when θ_1 and θ_2 were uncorrelated with the compensatory data, $r_{\hat{\theta}\bar{\theta}}$ showed a tendency to be larger than $r_{\hat{\theta}\theta_1}$ and $r_{\hat{\theta}\theta_2}$, and the gap between $r_{\hat{\theta}\bar{\theta}}$ and $r_{\hat{\theta}\theta_1}$ began to widen as $\rho_{\theta_1 \theta_2}$ approached the lower end of its scale. Across MIRT models,

$r_{\hat{\theta}\theta_2}$ decreased conspicuously when $\rho_{\theta_1\theta_2}$ changed from .45 to 0. As alternative measures of the interdimensional ability association, the observed Spearman rho correlation coefficients revealed similar numerical characteristics. On the whole, the $\hat{\theta}$ values could be thought of as an overestimate of the θ_1, θ_2, and $\bar{\theta}$ values to varying degrees. With the compensatory data, this type of bias fluctuated in magnitude across different levels of $\rho_{\theta_1\theta_2}$. With the noncompensatory data, the bias grew with the decrease of $\rho_{\theta_1\theta_2}$, and the growth rate was somewhat noteworthy when $\rho_{\theta_1\theta_2}$ dropped from .45 to 0. In most cases, $Bias_1$ was the smallest, and $Bias_2$ the largest. The *RMSE*s rose more rapidly than *ASE* as $\rho_{\theta_1\theta_2}$ decreased, and the increment of every error measure was more appreciable when $\rho_{\theta_1\theta_2}$ decreased from .45 to 0. $RMSE_1$ and $RMSE_3$ were each markedly smaller than $RMSE_2$, and $RMSE_3$ was the smallest of the three except when $\rho_{\theta_1\theta_2}$ equaled 0 with the compensatory data. As $\rho_{\theta_1\theta_2}$ decreased, more visible increase of *ANI* occurred with the noncompensatory data than with the compensatory data. This upward movement was noticeable with the noncompensatory data when $\rho_{\theta_1\theta_2}$ changed from .45 to 0. Except for the correlation coefficients, evaluative indices associated with $\rho_{\theta_1\theta_2}$ appeared larger with the noncompensatory data than with the compensatory data.

METHOD Effect

Table 7 provides evaluative indices to assess the CAT simulation results as a function of METHOD. In general, the $\hat{\theta}$ values were highly correlated with the θ_1 and $\bar{\theta}$ values. Although the observed Pearson product-moment correlation coefficients favored the Bayesian estimation methods slightly with the noncompensatory data, they increased almost negligibly as METHOD changed from maximum likelihood to Bayes EAP across MIRT models. With every ability estimation method, $r_{\hat{\theta}\bar{\theta}}$ always remained the largest, and $r_{\hat{\theta}\theta_2}$ the smallest. The observed Spearman rho correlation coefficients did not exactly match their Pearson product-moment counterparts. However, the fidelity pattern was basically retained. With the compensatory data, there was some indication that the bias of the CAT simulation results changed its direction from the positive to the negative as METHOD switched from maximum likelihood to Bayes EAP. No bias measures seemed to be a cause for particular concern. The maximum likelihood and the Bayes EAP estimation methods stood at each extreme of the bias scale, whereas the Bayesian modal estimation method produced the least amount of $Bias_1$ and $Bias_3$ in absolute value. With the noncompensatory data, the CAT simulation results overestimated θ_1, θ_2, and $\bar{\theta}$ to some degree. The bias measures decreased as METHOD switched from maximum likelihood to Bayesian sequential. These measures reached their highest levels with Bayesian modal and lowest levels with Bayes EAP. Across ability estimation methods for both MIRT models, $Bias_1$ invariably remained the smallest, and $Bias_2$ the largest. The *RMSE*s and *ASE* decreased when Bayesian sequential superseded maximum likelihood as the ability estimation method. These two evaluative indices tended to fluctuate with the subsequent use of the other Bayesian estimation methods. Generally speaking, $RMSE_2$ was twice as large as $RMSE_1$ and $RMSE_3$, and $RMSE_3$ was the smallest of the three. The Bayesian modal estimation method produced the lowest *ASE*. *ANI* varied across ability estimation methods and MIRT models. With the compensatory data,

the Bayesian sequential estimation method achieved the highest test efficiency, whereas the maximum likelihood estimation method exhibited the lowest. With the noncompensatory data, the Bayesian sequential estimation method lagged behind the other estimation methods in test efficiency, and the Bayesian modal estimation method was noted for its smallest *ANI* values. Except for the correlation coefficients, evaluative indices associated with METHOD appeared larger with the noncompensatory data than with the compensatory data.

Interaction Effect

As most of the two-dimensional evaluative indices showed that $\bar{\theta}$ is the type of ability estimate CAT can best produce, measures associated with θ_1 and θ_2 are excluded from this subsection to provide focus for the presentation.

Figures 2 and 3 depict an interaction between $\rho_{\theta_1\theta_2}$ and METHOD with $RMSE_3$ as the dependent variable for both compensatory and noncompensatory MIRT models. As it can be seen, the maximum likelihood estimation method generally yields more root mean square error than the Bayesian estimation methods so far as the difference between $\bar{\theta}$ and $\hat{\theta}$ is concerned. However, a lack of parallelism between their respective lines shows two distinct patterns of disparity between the two types of estimation methods across MIRT models: the disparity is less appreciable with the lower $\rho_{\theta_1\theta_2}$ levels than with the higher $\rho_{\theta_1\theta_2}$ levels based on the compensatory data (see Figure 2), and vice versa when the noncompensatory data are involved (see Figure 3). Figures 4 and 5 illustrate other forms of the $\rho_{\theta_1\theta_2}$-METHOD interaction. Based on the compensatory model, Figure 4 delineates some noticeable difference between the maximum likelihood estimation method and every Bayesian estimation method in *ASE* at higher $\rho_{\theta_1\theta_2}$ levels. This difference tapers off almost to a point when $\rho_{\theta_1\theta_2}$ approaches 0. Based on the noncompensatory model, Figure 5 shows an increase of *ANI* as $\rho_{\theta_1\theta_2}$ decreases. There is a sizable disparity between the Bayesian sequential and the Bayesian modal estimation methods when θ_1 and θ_2 are highly correlated. However, this disparity virtually vanishes as $\rho_{\theta_1\theta_2}$ reaches the lower end of its scale.

As a follow-up of the investigation of the first-order interaction, second-order interactions among $\rho_{\theta_1\theta_2}$, METHOD, and $\rho_{b_1b_2}$ were explored with $RMSE_3$, *ASE* (for the compensatory data) and *ANI* (for the noncompensatory data) as the dependent variable respectively. With the compensatory data, the $\rho_{\theta_1\theta_2}$-METHOD interaction behaved similarly across different levels of $\rho_{b_1b_2}$ in both cases; with the noncompensatory data, the $\rho_{\theta_1\theta_2}$-METHOD interaction had no differential effect across all levels of $\rho_{b_1b_2}$ in each situation.

Magnitudes of Factorial Effects

Table 8 contains measures of the magnitudes of the effects of the simulation factors across MIRT models. T1, T2, M1, and T1M1 are the four factorial effects associated with relatively large squared Pearson product-moment correlation coefficients based on the compensatory data.

Representing the linear and the quadratic trends, T1 (r^2 = .6952) and T2 (r^2 = .0562) altogether accounted for 75% of the variance in $LRMSE_3$. Their concurrent presence at $\rho_{\theta_1\theta_2}$ is highlighted in Figure 6, where the systematic increase of $LRMSE_3$ as $\rho_{\theta_1\theta_2}$ decreases indicates the linear trend, and the modest acceleration of this increase at the $\rho_{\theta_1\theta_2}$ level of .45 indicates the quadratic trend. Contrasting the maximum likelihood estimation method with all Bayesian estimation methods, M1 (r^2 = .1839) was related to 18% of the variance in $LRMSE_3$. Figure 7 displays this distinction in $LRMSE_3$ between the two types of estimation methods. Finally, T1M1 (r^2 = .0236) stands for the interaction between T1 and M1, and only explained 2% of the variance in $LRMSE_3$. Largely a replica of Figure 2, the figure associated with this factorial effect is omitted for the sake of brevity.

The aforementioned factorial effects warrant our special attention in that they accounted for 96% of the variance in $LRMSE_3$ based on the compensatory MIRT model. Despite the presence of the $\rho_{\theta_1\theta_2} - \text{METHOD}$ interaction, it is legitimate to discuss main effects in this situation for three reasons. First, the interaction was quantitatively small ($r^2 = .0236$). Second, the main effects were obvious in that regardless of the observed interaction, higher $LRMSE_3$ values were consistently attributable to lower $\rho_{\theta_1\theta_2}$ levels and the maximum likelihood estimation method. Third, the main effects and their interaction were independent of one another as they were coded in orthogonal contrasts.

The factorial effects T1, T2, and M1 are associated with relatively large squared Pearson product-moment correlation coefficients based on the noncompensatory data.. A combination of the linear trend T1 (r^2 = . 6006) and the quadratic trend T2 (r^2 = . 0301) explained 63% of the variance in $LRMSE_3$. Their dual effect on $\rho_{\theta_1\theta_2}$ can be seen in Figure 8, where the continuous increase of $LRMSE_3$ over the decreasing $\rho_{\theta_1\theta_2}$ indicates the linear trend, and the gradual acceleration of this increase at the $\rho_{\theta_1\theta_2}$ level of .45 indicates the quadratic trend. By directly comparing the maximum likelihood estimation method with all Bayesian estimation methods, M1 (r^2 = .3360) accounted for 34% of the variance in $LRMSE_3$. Figure 9 illustrates this distinction in $LRMSE_3$ between the two types of estimation methods.

As the aforementioned factorial effects were associated with 97% of the variance in $LRMSE_3$, they are of major importance to the CAT simulation results based on the noncompensatory MIRT model. Figures 3 and 5 have provided some evidence of the presence of a $\rho_{\theta_1\theta_2} - \text{METHOD}$ interaction. However, this effect turned out to be trivial because T1M1, T1M3, and T2M1, the only representations of the $\rho_{\theta_1\theta_2} - \text{METHOD}$ interaction in Table 8 that carries a nonzero r^2 value, explained only 0.3% of the variance in the dependent variable.

Unidimensional versus Two-dimensional Data

Table 9 reports the average evaluative indices of CAT simulation results across different ability estimation methods when the item response matrices were based on the unidimensional IRT model and the two-dimensional compensatory and noncompensatory MIRT models respectively. The equally weighted average of θ_1 and θ_2 was treated as the true ability when the average evaluative indices based on the two-dimensional MIRT model were computed.

CAT simulations involving the unidimensional data brought slightly better evaluative indices than those involving the two-dimensional compensatory data did in many respects. They were associated with more desirable evaluative indices than those involving the two-dimensional noncompensatory data across all ability estimation methods. Such differences were consistent in fidelity measures, and less stable with bias measures. *RMSE* showed a more substantial improvement when the unidimensional item response matrices replaced their two-dimensional counterparts for CAT simulations, and this change was especially conspicuous with the maximum likelihood estimation method. It was interesting that the smallest *ASE* values were related to the two-dimensional compensatory data except when the maximum likelihood estimation method was employed. On average, CAT simulations based on the unidimensional data required more items per testing session than those based on the two-dimensional compensatory data did by 3%. These unidimensional data-based CAT simulations, in turn, required much fewer items per testing session than those based on the two-dimensional noncompensatory data did by 32%.

The differences between the unidimensional and the two-dimensional compensatory data with respect to the CAT simulation results manifested themselves somewhat differently if the correlation between θ_1 and θ_2 was taken into consideration as well. Table 10 presents such differences with the $\rho_{\theta_1\theta_2}$ value set at .89. It is clear that CAT simulation results based on the two-dimensional compensatory data could be more accurate than those based on the unidimensional data so long as (a) the Bayesian estimation methods were applied, (b) the θ_1 and θ_2 values were highly correlated, and (c) $\bar{\theta}$ was taken as the true ability. *Bias* was the only evaluative index that failed to follow this pattern when Bayesian sequential or Bayes EAP was chosen as the ability estimation method.

Discussion

This Monte Carlo study investigated the robustness of the unidimensional 3PL IRT model to the violation of its unidimensionality assumption in CAT applications. The correlation coefficients and RMSEs derived from the study indicated that, for both compensatory and noncompensatory MIRT models, the unidimensionally derived CAT ability estimates approximate the average of the true two-dimensional abilities. This finding is consistent with what Ansley and Forsyth (1985), Way et al. (1988), and De Ayala (1992) discovered in their respective studies concerning the relationship between the unidimensional IRT-based ability estimates and the average of the true two-dimensional abilities.

The results of this study also provided answers to the research questions previously specified:

1. The decrease in the interdimensional item difficulty correlation has little effect on CAT simulation results despite some slightly unfavorable changes it brings to the evaluative measures of the ability estimates. De Ayala also noticed this phenomenon in his 1992 study regarding the influence of dimensionality on CAT ability estimation.

2. Except when the bias measures derived from the compensatory data are involved, the decrease in the interdimensional ability correlation is associated with increasingly undesirable evaluative measures of the CAT simulation results. This is especially so when $\rho_{\theta_1\theta_2}$ falls from .45 to 0.

3. The application of different ability estimation methods affects the CAT simulation results differently, and the Bayesian estimation methods almost always produce more accurate ability estimates than does the maximum likelihood estimation method. With the compensatory data, the Bayesian modal estimation method suffers the least bias and error, and the Bayesian sequential estimation method proves most test-efficient. With the noncompensatory data, on the other hand, the Bayes EAP estimation method produces the least bias, the Bayesian sequential estimation method yields the smallest RMSE, and the Bayesian modal estimation method provides the most acceptable *ASE* and *ANI* measures. It should be noted that across the MIRT models the differences between the Bayesian modal and the Bayes EAP estimation methods are no greater than 0.04 in fidelity, bias, and error measures.

4. The interaction between the interdimensional ability correlation and the ability estimation method slightly affects the CAT simulation results based on the compensatory data. As $\rho_{\theta_1\theta_2}$ decreases, the differences in error measures tend to diminish between the maximum likelihood estimation method and every Bayesian estimation method. A slightly different interaction can be found with the CAT simulation results associated with the noncompensatory data. However, this effect turns out to be inconsequential.

The interdimensional ability correlation and the ability estimation method have large effects on the CAT simulation results because of the considerable amount of variance in the dependent variable $LRMSE_3$ they can account for. With the compensatory and the noncompensatory data alike, the effects of these two simulation factors are mainly reflected in the linear and the quadratic trends associated with $\rho_{\theta_1\theta_2}$, and the differences between the maximum likelihood estimation method and every Bayesian estimation method.

5. Except when *ASE* and *ANI* are related to the compensatory data, CAT simulations based on unidimensional data usually produce more desirable results than those based on two-dimensional data, assuming that $\bar{\theta}$ can be the treated as the true ability underlying the two-dimensional data. This phenomenon is especially conspicuous with the noncompensatory data-based CAT simulations. However, CAT simulations associated with the two-dimensional compensatory data can outperform those associated with the unidimensional data on most evaluative indices when θ_1 and θ_2 are highly correlated, and the Bayesian estimation methods are applied.

Although it is difficult to draw definitive conclusions about the robustness of the unidimensional 3PL IRT model in CAT applications based on the results of a single Monte Carlo study, tentative explanations can be made regarding some of the findings summarized above.

First, it is possible that the close association between the unidimensionally derived ability estimate and the average of θ_1 and θ_2 is merely a function of the MIRT model employed for data generation. An inspection of the two-dimensional version of the compensatory and the noncompensatory MIRT models indicates that no matter how dominant θ_1 is, these models weigh θ_1 and θ_2 equally when the probability of a correct item response is computed. As a result, when the model-based two-dimensional ability parameters are unidimensionally estimated in the CAT simulation process, the observed estimate may carry an equal amount of information about both θ_1 and θ_2. It is not clear whether the unidimensional calibration of the two-dimensional item parameters prior to a CAT simulation plays a role here as well. Second, the reason why the interdimensional item difficulty correlation has little effect on CAT simulation results may be found in the testing procedure itself. A given $\rho_{b_1 b_2}$ level is a measure of the association between the b_1 and b_2 values of an item pool. Although the numerical change of this index may affect the difficulty of a test (Ansley & Forsyth, 1985), it cannot alter the ability estimate in a CAT setting if the item pool covers the whole difficulty continuum adequately, and the test is of reasonable length. Third, the positive effect of the high interdimensional ability correlation on CAT simulation results may stem from a dimensional change of the corresponding item response data set. It is true that even when the vectors are perfectly correlated, an MIRT model is different from a unidimensional IRT model (Ansley & Forsyth, 1985). Yet, a high $\rho_{\theta_1 \theta_2}$ level can enhance the model-data fit by rendering the data set nearly unidimensional, and eventually improve the ability estimation.

This Monte Carlo study was limited in several respects. First, because the study only employed the two-dimensional version of the compensatory and the noncompensatory MIRT models for data generation with the item and ability parameters manipulated to exhibit specific characteristics, its results were not universally generalizable. Second, the CAT simulation procedure adopted in the study included neither content constraints nor item exposure control. As a result, the simulated CAT sessions might not truthfully reflect the reality. Third, the simulees' abilities in the study were generated to approximate a bivariate normal distribution with zero mean and unit variance for each ability dimension. Consequently, valid conditional evaluative indices were difficult to obtain at both extremes of the ability continuum to determine if multidimensionality affected CAT simulations equally across the whole range of abilities.

Despite its limitations, this Monte Carlo study has some implications for testing practitioners. First, when planning to use unidimensional IRT-based CAT, it is advisable to exercise caution if the unidimensionality assumption is likely to be violated. In an ideal world, an accurate ability estimate derived from multidimensional data should reflect the relative potency of individual ability dimensions. Yet this is probably unobtainable considering the results of this and other Monte Carlo studies. In the final analysis, it may not be practical after all to expect every dimension to be proportionally represented in a unidimensional ability estimate, because the unidimensional IRT model itself has no built-in mechanism to reconstruct the dimensional pattern of the multidimensional input data. What may be practical instead is to identify the real world circumstances under which a composite score of a linear combination of the dimensions is acceptable to the test user, and a slight departure from unidimensionality will not compromise the utility of CAT results. De Ayala (1992) found that, in a classroom test setting, the instructor may be only interested in rank ordering the students on their subject-related overall ability rather than on the separate dimensions of this ability. However, given that the unidimensional 3PL IRT

model is robust to the violation of its unidimensionality assumption only under limited conditions, preparatory work must be done prior to a CAT application so as to ensure that the abilities to be tested are highly correlated in the target examinee population, that compensation involving unbalanced ability levels on different dimensions can occur, and that a Bayesian estimation method will be employed to score the test.

Second, when unidimensional IRT-based CAT has been deemed suitable for an examinee population in a given testing situation, it would be beneficial to find out which ability estimation method can provide the most accurate scores possible. Because the increase in interdimensional item difficulty correlation hardly improves the measurement quality, the selection of an appropriate scoring method becomes one of the few decisions testing practitioners can make to ensure the quality of a CAT administration. Of the four ability estimation methods examined in this Monte Carlo study, the maximum likelihood estimation method may not be preferred due to its relatively poor performance. Among the Bayesian estimation methods, Bayesian sequential may also need to be ruled out because its assumption of a normal posterior distribution before each update is considered invalid, and its ability estimate fluctuates as a function of the presentation order of the items (Bock & Mislevy, 1982; Thissen & Mislevy, 1990). It may be a little difficult to choose between the Bayesian modal and the Bayes EAP estimation methods. Although the former is computationally less efficient than the latter, each has its own merits. Therefore, considering the fact that the differences between the two Bayesian estimation methods are minor in many instances, testing practitioners can select either of them depending on the type of evaluative indices they are interested in. Wang and Vispoel (1998) found the Bayes EAP estimation method superior to other Bayesian estimation methods with unidimensional data. This superiority was not substantiated in this Monte Carlo study.

A number of directions for future research can be proposed based on the results of this CAT-related model robustness study. First, the number of dimensions needs to be increased when MIRT models are used for data generation, because live-testing situations may involve more dimensions. Second, more realistic item parameters need to be incorporated into the simulation process. Considering the fact that most of the standardized tests are developed according to very detailed test specifications, researchers should embed characteristics of such tests in their simulated item pools just as Ansley and Forsyth (1985), Davey, Nering, and Thompson (1997); and Wang and Vispoel (1998) did. Third, a greater variety of test conditions need to be modeled in CAT simulations. Improved understanding of the CAT procedures will be gained when researchers begin to manipulate more factors in the testing algorithm.

Yoes (1993) suggested that a set of standard test conditions (e.g., sample size, dimension, and test length) need to be incorporated into future IRT-based Monte Carlo studies so that research findings can become more comparable. The measurement community will benefit from following his suggestion.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.

Ansley, T. N., & Forsyth, R. A. (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: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitken, 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.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series 97-4). Iowa City, IA: American College Testing.

De Ayala, R. J. (1992). The influence of dimensionality on CAT ability estimation. *Educational and Psychological Measurement, 52*, 513-528.

Doody, E. N. (1985, March). *Examining the effects of multidimensional data on ability and item parameter estimation using the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Doody-Bogan, E., & Yen, W. M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13,* 373-389.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hattie, J. (1981). *Decision criteria for determining unidimensionality*. Unpublished doctoral dissertation. University of Toronto, Canada.

Hoffman, P. J. (1959). Generating variables with arbitrary properties. *Psychometrika*, *24*, 265-267.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR82-1). Iowa City, IA: American College Testing.

MicroCAT Testing System (Version 3.0) [Computer program]. (1993). St. Paul, MN: Assessment Systems Corporation.

Mislevy, R. J., & Bock, R. D. (1982). BILOG, maximum likelihood item analysis and test scoring: Logistic model [Computer program]. Mooresville, IN: Scientific Software.

Moonan, W. J. (1957). Linear transformation to a set of stochastically dependent normal variables. *American Statistical Association Journal 52*, 247-252.

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.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, No. 17*.

SAS/IML (Version 6) [Computer programming language] (1990). Cary, NC: SAS Institute.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103-135). Hillsdale, NJ: Lawrence Erlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Vale, C. D., & Gialluca, K. A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters* (Research Report No. ONR85-4). St. Paul, MN: Assessment Systems Corporation.

Wainer, H. (1990). Introduction and history. n H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 1-21). Hillsdale, NJ: Lawrence Erlbaum.

Wang, T., & Vispoel, W. P., 1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12,* 239-252.

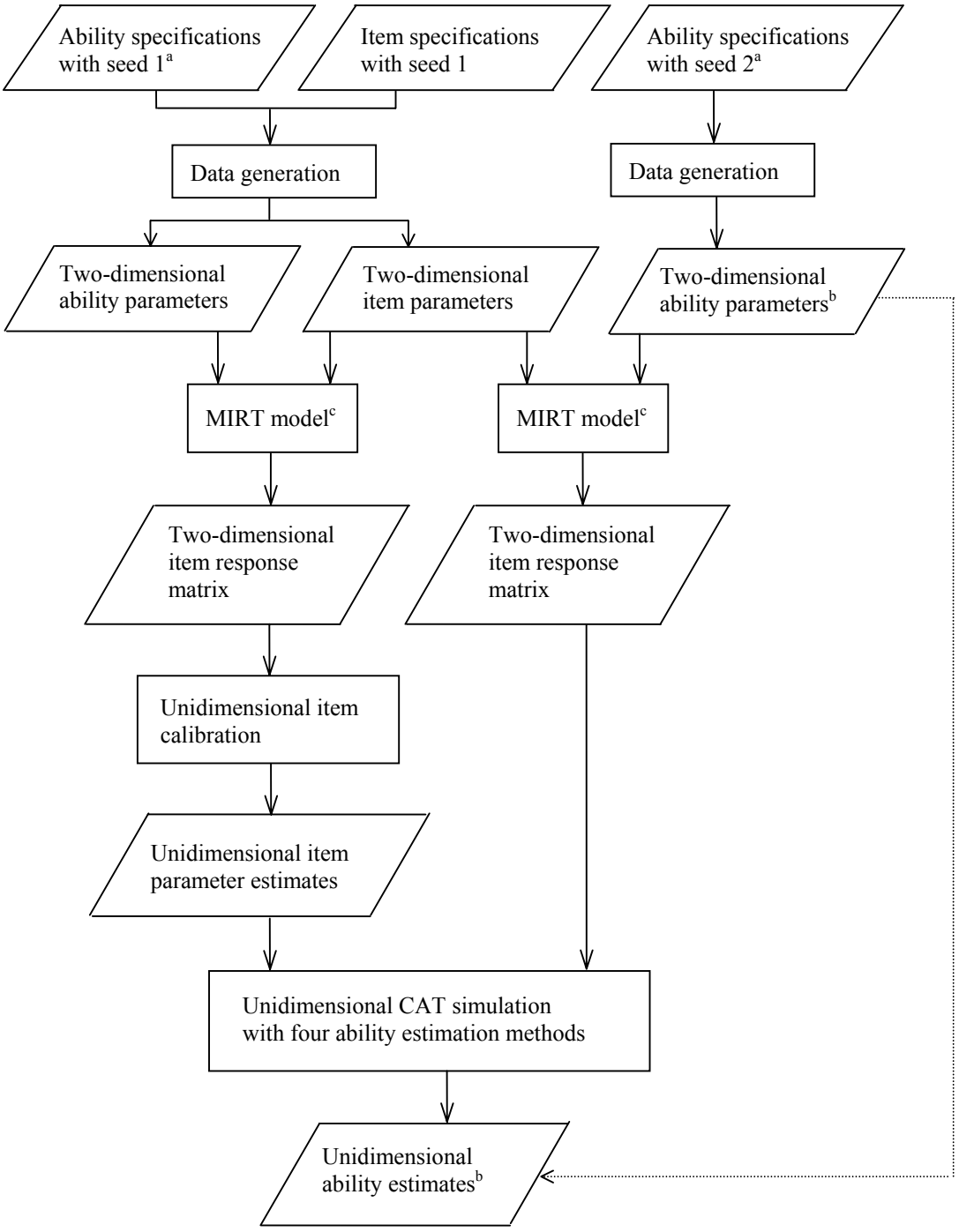
Weiss, D. J., & Suhadolnik, D. (1982). Robustness of adaptive testing to multidimensionality. In D. J. Weiss (Ed.), *Item Response Theory and Computerized Adaptive Testing Conference proceedings* (pp. 248-280). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 69-95). Boston: Kluwer.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum RM-76-6). Princeton, NJ: Educational Testing Service.

Yoes, M. E. (1993). *A comparison of the effectiveness of item parameter estimation techniques used with the 3-parameter logistic item response theory model*. Unpublished doctoral dissertation. The University of Minnesota.

Figure 1
Basic Structure of a Two-dimensional Data-based CAT simulation



Note: [a]Although seed values were different, ability specifications in the simulation remained the same.
[b]The disparities between the two-dimensional ability parameters and their unidimensional estimates are the focus of this dissertation study.
[c]Both compensatory and noncompensatory MIRT models were used.

Table 1

Summary Statistics for the Generated Two-dimensional and Unidimensional Item Discrimination and Difficulty Parameters

Parameter	*M*	*SD*	*Min.*	*Max.*
Two-dimensional[a]				
a_1	1.23	0.44	0.51	2.00
a_2	0.63	0.21	0.25	1.00
Unidimensional				
a	1.25	0.43	0.51	1.99
Two-dimensional				
		$r_{b_1b_2} = .10$ (.07)[b]		
b_1	0.00	1.19	-2.00	2.00
b_2	-1.00	0.40	-1.74	-0.31
		$r_{b_1b_2} = .45$ (.43)		
b_1	0.00	1.19	-2.00	2.00
b_2	-1.00	0.40	-1.86	-0.17
		$r_{b_1b_2} = .63$ (.62)		
b_1	0.00	1.19	-2.00	2.00
b_2	-1.00	0.40	-1.90	-0.12
		$r_{b_1b_2} = .77$ (.76)		
b_1	0.00	1.19	-2.00	2.00
b_2	-1.00	0.40	-1.90	-0.11
		$r_{b_1b_2} = .89$ (.89)		
b_1	0.00	1.19	-2.00	2.00
b_2	-1.00	0.40	-1.87	-0.14
Unidimensional				
b	0.00	1.19	-2.00	2.00

Note.

[a] $r_{a_1a_2} = -.02$; $a_1 : a_2 = 1.97$.

[b] The value in parentheses is the observed $r_{b_1b_2}$.

Table 2

Summary Statistics for the Generated Two-dimensional and Unidimensional Ability Parameters for Item Calibration and CAT Simulations

	Item calibration				CAT simulations			
Parameter	*M*	*SD*	*Min.*	*Max.*	*M*	*SD*	*Min.*	*Max.*
Two-dimensional								
				$r_{\theta_1\theta_2} = .00$ (.01, .03)				
θ_1	0.01	0.99	-3.35	3.18	0.01	1.00	-3.07	3.21
θ_2	0.01	1.00	-3.33	3.41	0.00	1.00	-3.10	3.60
				$r_{\theta_1\theta_2} = .45$ (.45, .47)				
θ_1	0.01	0.99	-3.35	3.18	0.01	1.00	-3.07	3.21
θ_2	0.02	1.00	-3.28	3.48	0.00	1.01	-3.23	3.77
				$r_{\theta_1\theta_2} = .63$ (.63, .64)				
θ_1	0.01	0.99	-3.35	3.18	0.01	1.00	-3.07	3.21
θ_2	0.01	1.00	-3.33	3.40	0.01	1.01	-3.39	3.69
				$r_{\theta_1\theta_2} = .77$ (.77, .78)				
θ_1	0.01	0.99	-3.35	3.18	0.01	1.00	-3.07	3.21
θ_2	0.01	1.00	-3.38	3.31	0.01	1.02	-3.45	3.66
				$r_{\theta_1\theta_2} = .89$ (.89, .89)				
θ_1	0.01	0.99	-3.35	3.18	0.01	1.00	-3.07	3.21
θ_2	0.01	0.99	-3.30	3.27	0.01	1.02	-3.41	3.60
Unidimensional								
θ	0.01	1.00	-3.28	3.40	0.00	1.00	-3.25	3.18

Table 3

Nonlinear Factor Analysis of Representative Item Response Matrices Based on the Generated Two-dimensional Item and Ability Parameters

Item response matrix	Factor solution	RMSR	RMSR reduction (%)
baco1000 (2,000 × 123)	1-factor	0.0047	
	2-factor	0.0036	23%
	3-factor	0.0030	17%
r1nc1000 (2,000 × 123)	1-factor	0.0054	
	2-factor	0.0044	19%
	3-factor	0.0043	3%
r2co4545 (2,000 × 123)	1-factor	0.0041	
	2-factor	0.0031	24%
	3-factor	0.0029	5%
r3nc6363 (1,000 × 123)	1-factor	0.0063	
	2-factor	0.0059	6%
	3-factor	0.0058	2%
r4co7777 (1,000 × 123)	1-factor	0.0060	
	2-factor	0.0044	26%
	3-factor	0.0043	2%
r5nc8989 (1,000 × 123)	1-factor	0.0062	
	2-factor	0.0057	8%
	3-factor	0.0056	2%

Note. The names of the item response matrices contain the following information (with the decimal points excluded for clarity of presentation where applicable):

1. Research design: basic research design (ba) or replication (r1-r5).
2. Simulation model: co (compensatory) or nc (noncompensatory).
3. $\rho_{b_1 b_2}$: 01, 45, 63, 77, or 89.
4. $\rho_{\theta_1 \theta_2}$: 00, 45, 63, 77, or 89.

The first three matrices were generated for item calibration, whereas the remaining three for CAT simulations.

Table 4

Differences Between the SimuCAT and the MicroCAT Simulation Results

Simulee	Maximum likelihood		Bayesian sequential		Bayesian modal		Bayes EAP	
	$DIF_{\hat{\theta}}$	$DIF_{SD^2_{\hat{\theta}}}$	$DIF_{\hat{\theta}}$	$DIF_{SD^2_{\hat{\theta}}}$	$DIF_{\hat{\theta}}$	$DIF_{SD^2_{\hat{\theta}}}$	$DIF_{\hat{\theta}}$	$DIF_{SD^2_{\hat{\theta}}}$
2	-0.12	-0.02						
17	0.06							
19	0.03							
40	0.20	0.01						
42					0.05			
57					0.06			
65	-0.08	0.02						
69	-0.08							
72					-0.10			
78	-0.06							
85	-0.02							
88	-0.07							
93	-0.06	0.01						

Note. $DIF_{\hat{\theta}}$ and $DIF_{SD^2_{\hat{\theta}}}$ are the SimuCAT and MicroCAT differences in terms of $\hat{\theta}s$ and $SD_{\hat{\theta}}{}^2$s. This table excludes simulees with absolute difference values less than 0.005 in both $DIF_{\hat{\theta}}$ and $DIF_{SD^2_{\hat{\theta}}}$ across the four ability estimation methods. With the simulees listed in the table, absolute difference values less than 0.005 in $DIF_{\hat{\theta}}$ or $DIF_{SD^2_{\hat{\theta}}}$ are not presented for clarity.

Table 5

Average Evaluative Indices of CAT Simulation Results by $\rho_{b_1 b_2}$ *and MIRT Model*

Evaluative index	$\rho_{b_1 b_2}$				
	.10	.45	.63	.77	.89
			Fidelity		
$r_{\hat{\theta}\theta_1}$	0.941 (0.900)	0.942 (0.902)	0.942 (0.902)	0.943 (0.902)	0.943 (0.901)
$r_{\hat{\theta}\theta_2}$	0.732 (0.724)	0.732 (0.724)	0.733 (0.725)	0.733 (0.726)	0.733 (0.728)
$r_{\hat{\theta}\bar{\theta}}$	0.949 (0.919)	0.949 (0.921)	0.950 (0.921)	0.950 (0.922)	0.950 (0.922)
$rs_{\hat{\theta}\theta_1}$	0.942 (0.903)	0.943 (0.904)	0.943 (0.904)	0.943 (0.903)	0.943 (0.903)
$rs_{\hat{\theta}\theta_2}$	0.721 (0.713)	0.720 (0.713)	0.720 (0.713)	0.719 (0.714)	0.719 (0.715)
$rs_{\hat{\theta}\bar{\theta}}$	0.951 (0.923)	0.950 (0.923)	0.950 (0.923)	0.949 (0.923)	0.949 (0.923)
			Bias		
$Bias_1$	0.004 (0.022)	0.001 (0.021)	0.001 (0.020)	0.000 (0.018)	-0.001 (0.017)
$Bias_2$	0.011 (0.029)	0.008 (0.028)	0.008 (0.027)	0.007 (0.025)	0.006 (0.024)
$Bias_3$	0.007 (0.025)	0.005 (0.025)	0.005 (0.023)	0.003 (0.021)	0.003 (0.021)
			Error		
$RMSE_1$	0.346 (0.447)	0.343 (0.442)	0.342 (0.441)	0.340 (0.441)	0.338 (0.443)
$RMSE_2$	0.703 (0.728)	0.700 (0.726)	0.700 (0.725)	0.699 (0.723)	0.698 (0.721)
$RMSE_3$	0.331 (0.406)	0.327 (0.402)	0.326 (0.400)	0.325 (0.398)	0.323 (0.398)
ASE	0.202 (0.278)	0.200 (0.274)	0.199 (0.272)	0.198 (0.271)	0.197 (0.270)
			Test efficiency		
ANI	20.846 (28.458)	20.799 (28.426)	20.782 (28.404)	20.772 (28.380)	20.758 (28.356)

Note.
Evaluative indices based on the compensatory data are not parenthesized; evaluative indices based on the noncompensatory data are parenthesized.

Table 6

Average Evaluative Indices of CAT Simulation Results by $\rho_{\theta_1\theta_2}$ *and MIRT Model*

Evaluative index	$\rho_{\theta_1\theta_2}$				
	.00	.45	.63	.77	.89
			Fidelity		
$r_{\hat{\theta}\theta_1}$	0.905 (0.794)	0.933 (0.893)	0.947 (0.923)	0.958 (0.942)	0.968 (0.954)
$r_{\hat{\theta}\theta_2}$	0.385 (0.426)	0.693 (0.679)	0.794 (0.772)	0.866 (0.844)	0.924 (0.907)
$r_{\hat{\theta}\bar{\theta}}$	0.901 (0.851)	0.948 (0.917)	0.960 (0.934)	0.967 (0.947)	0.972 (0.956)
$rs_{\hat{\theta}\theta_1}$	0.900 (0.793)	0.932 (0.894)	0.948 (0.925)	0.961 (0.945)	0.973 (0.960)
$rs_{\hat{\theta}\theta_2}$	0.359 (0.404)	0.674 (0.662)	0.782 (0.760)	0.860 (0.837)	0.925 (0.906)
$rs_{\hat{\theta}\bar{\theta}}$	0.892 (0.850)	0.947 (0.918)	0.962 (0.937)	0.971 (0.950)	0.978 (0.961)
			Bias		
$Bias_1$	-0.002 (0.029)	-0.003 (0.023)	0.001 (0.019)	0.003 (0.015)	0.006 (0.011)
$Bias_2$	0.012 (0.043)	0.006 (0.032)	0.007 (0.026)	0.007 (0.019)	0.008 (0.013)
$Bias_3$	0.005 (0.036)	0.002 (0.028)	0.004 (0.023)	0.005 (0.017)	0.007 (0.012)
			Error		
$RMSE_1$	0.439 (0.657)	0.371 (0.478)	0.334 (0.408)	0.300 (0.355)	0.265 (0.315)
$RMSE_2$	1.114 (1.085)	0.796 (0.822)	0.656 (0.694)	0.533 (0.575)	0.403 (0.447)
$RMSE_3$	0.477 (0.558)	0.341 (0.424)	0.297 (0.375)	0.270 (0.338)	0.247 (0.308)
ASE	0.211 (0.306)	0.197 (0.277)	0.195 (0.268)	0.196 (0.261)	0.196 (0.255)
			Test efficiency		
ANI	21.052 (29.713)	20.737 (28.902)	20.726 (28.367)	20.725 (27.844)	20.717 (27.198)

Note.
Evaluative indices based on the compensatory data are not parenthesized; evaluative indices based on the noncompensatory data are parenthesized.

Table 7
Average Evaluative Indices of CAT Simulation Results by METHOD and MIRT Model

Evaluative index	METHOD			
	Maximum likelihood	Bayesian sequential	Bayesian modal	Bayes EAP
		Fidelity		
$r_{\hat{\theta}\theta_1}$	0.936 (0.892)	0.942 (0.905)	0.945 (0.905)	0.947 (0.903)
$r_{\hat{\theta}\theta_2}$	0.728 (0.713)	0.731 (0.728)	0.734 (0.727)	0.736 (0.734)
$r_{\hat{\theta}\bar{\theta}}$	0.944 (0.908)	0.948 (0.925)	0.952 (0.924)	0.954 (0.927)
$rs_{\hat{\theta}\theta_1}$	0.941 (0.902)	0.942 (0.904)	0.944 (0.905)	0.944 (0.902)
$rs_{\hat{\theta}\theta_2}$	0.719 (0.706)	0.718 (0.717)	0.721 (0.712)	0.721 (0.720)
$rs_{\hat{\theta}\bar{\theta}}$	0.949 (0.918)	0.949 (0.925)	0.951 (0.923)	0.951 (0.926)
		Bias		
$Bias_1$	0.013 (0.018)	0.003 (0.015)	0.000 (0.038)	-0.012 (0.006)
$Bias_2$	0.020 (0.025)	0.010 (0.022)	0.007 (0.045)	-0.005 (0.013)
$Bias_3$	0.016 (0.022)	0.006 (0.019)	0.003 (0.042)	-0.008 (0.010)
		Error		
$RMSE_1$	0.388 (0.519)	0.334 (0.415)	0.323 (0.416)	0.322 (0.421)
$RMSE_2$	0.741 (0.810)	0.688 (0.696)	0.684 (0.698)	0.688 (0.694)
$RMSE_3$	0.390 (0.515)	0.311 (0.361)	0.300 (0.364)	0.304 (0.363)
ASE	0.227 (0.302)	0.193 (0.264)	0.182 (0.258)	0.194 (0.269)
		Test efficiency		
ANI	21.002 (28.267)	20.569 (29.371)	20.667 (27.466)	20.927 (28.514)

Note.
Evaluative indices based on the compensatory data are not parenthesized; evaluative indices based on the noncompensatory data are parenthesized.

Figure 2

Effect of $\rho_{\theta 1 \theta 2}$-METHOD Interaction on *RMSE* $_3$ Based on the Compensatory MIRT Model

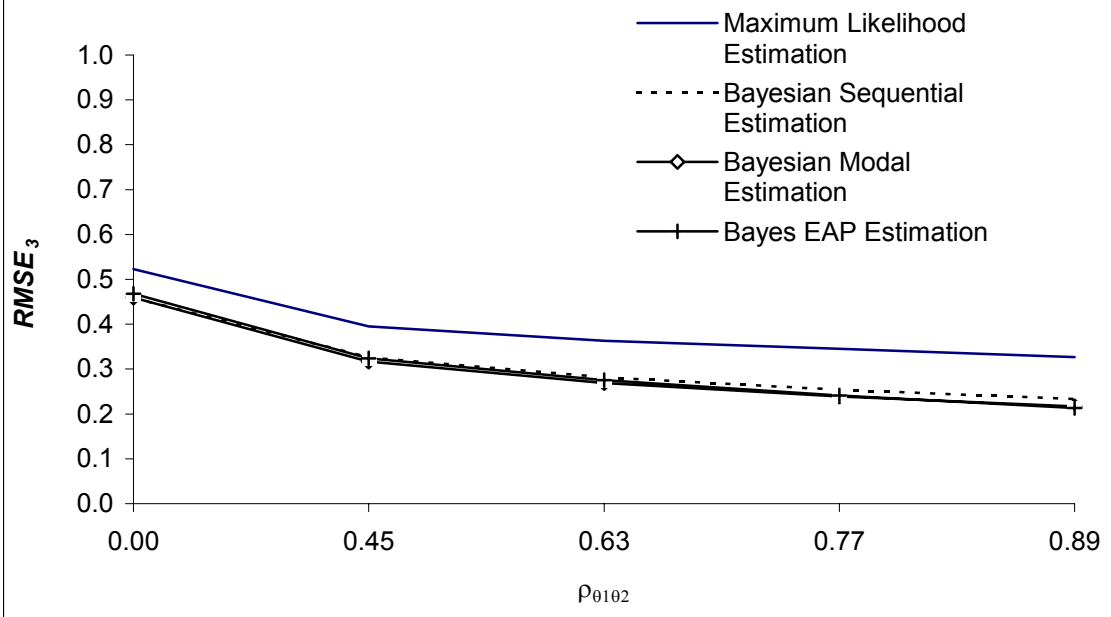


Figure 3

Effect of $\rho_{\theta 1 \theta 2}$-METHOD Interaction on *RMSE* $_3$ Based on the Noncompensatory MIRT Model

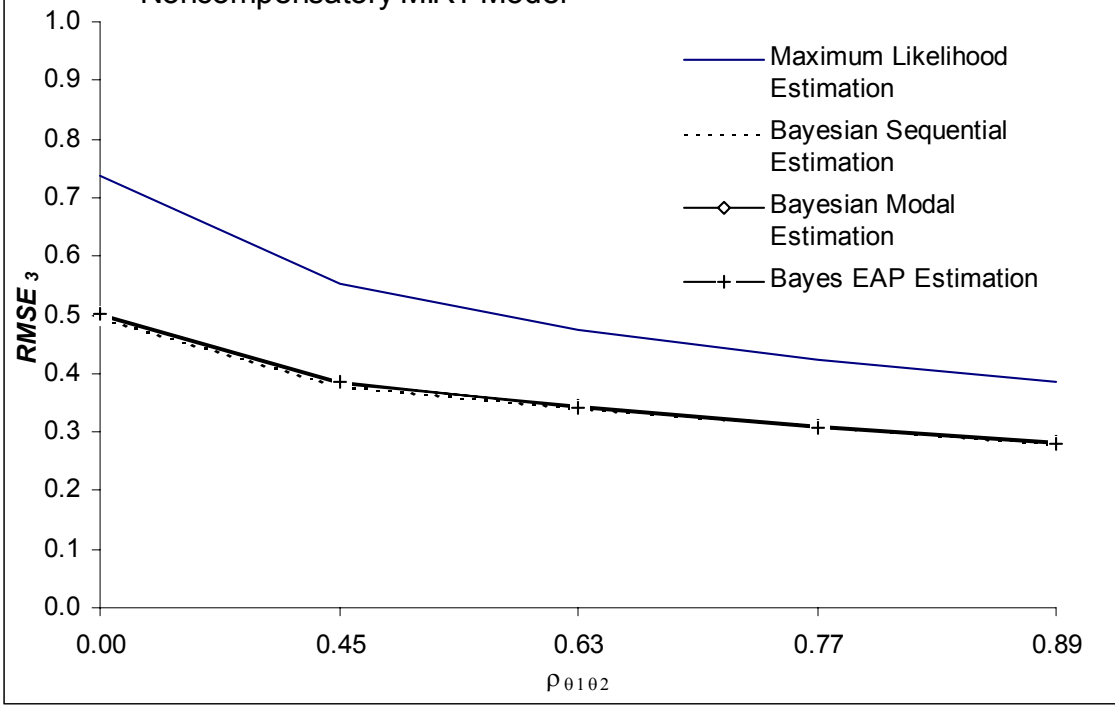


Figure 4

Effect of $\rho_{\theta_1\theta_2}$–METHOD Interaction on *ASE* Based on the Compensatory MIRT Model

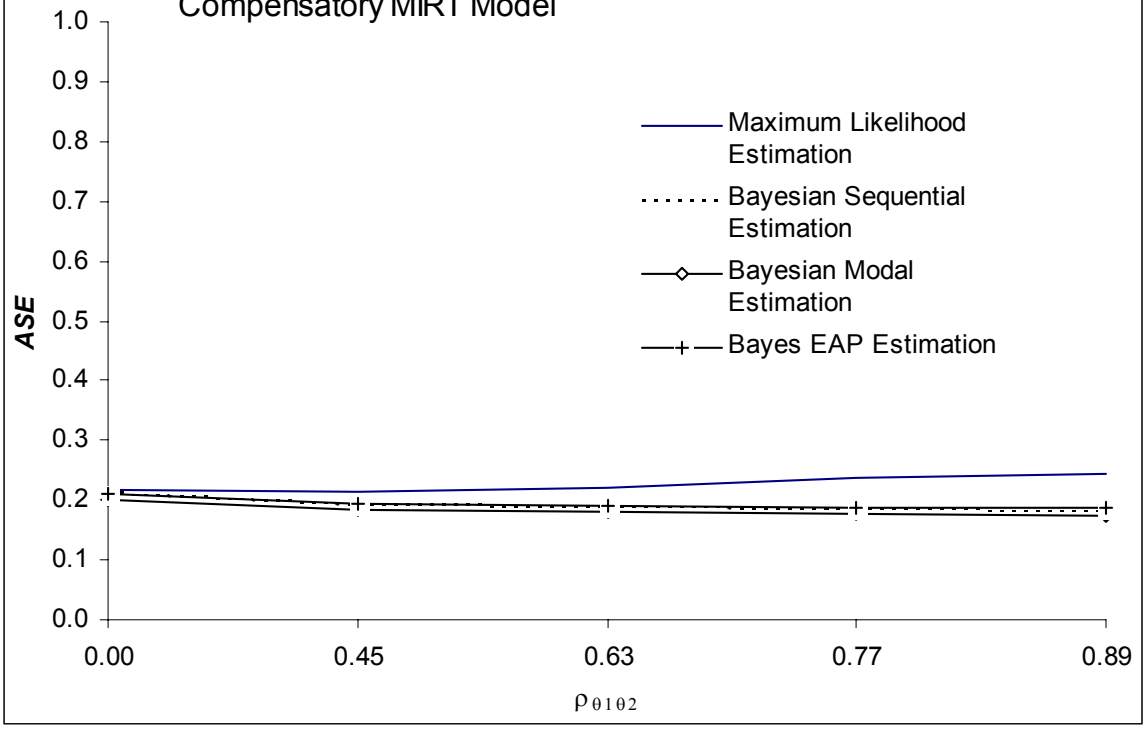


Figure 5

Effectct of $\rho_{\theta_1\theta_2}$–METHOD Interaction on *ANI* Based on the Noncompensatory MIRT Model

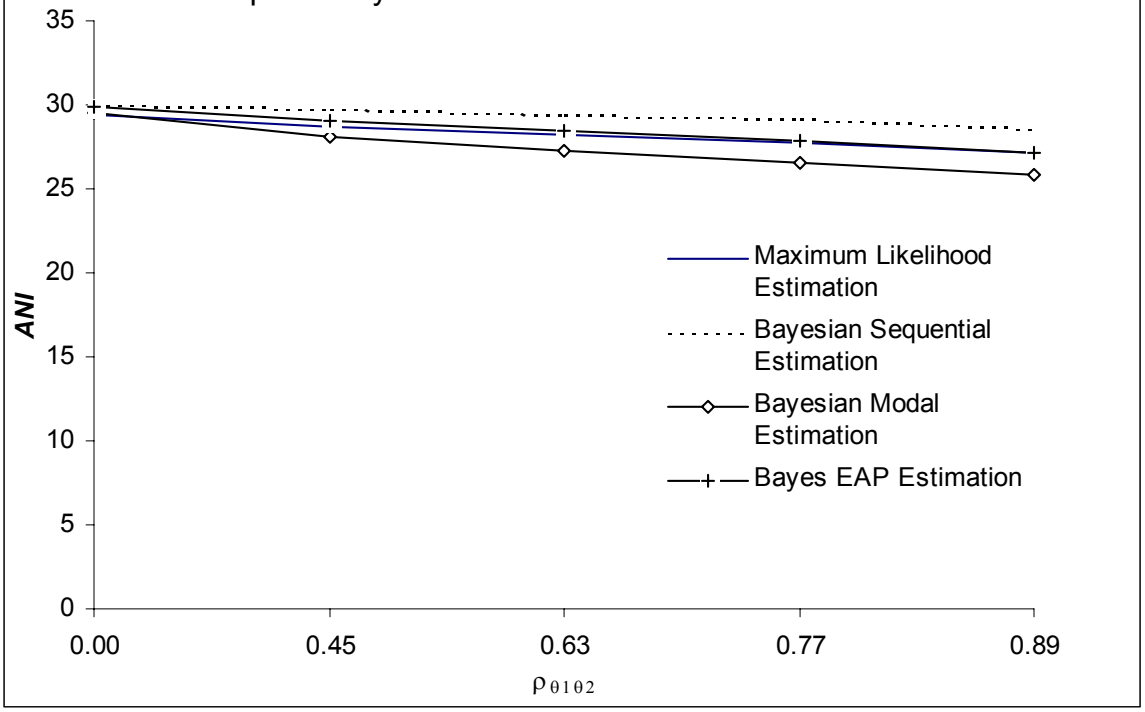


Table 8
Magnitudes of the Effects of the Simulation Factors by MIRT Model

Effect	r	r^2	Effect	r	r^2
B1	0.0299	0.0009	T1M2	-0.0472	0.0022
	(0.0305)	(0.0009)		(-0.0064)	(0.0000)
B2	0.0065	0.0000	T1M3	-0.0171	0.0003
	(0.0121)	(0.0001)		(-0.0125)	(0.0002)
T1	0.8338	0.6952	T2M1	-0.0125	0.0002
	(0.7750)	(0.6006)		(0.0171)	(0.0003)
T2	0.2371	0.0562	T2M2	-0.0007	0.0000
	(0.1736)	(0.0301)		(-0.0022)	(0.0000)
M1	0.4288	0.1839	T2M3	0.0113	0.0001
	(0.5796)	(0.3360)		(0.0015)	(0.0000)
M2	0.0554	0.0031	B1T1M1	-0.0107	0.0001
	(-0.0081)	(0.0001)		(0.0031)	(0.0000)
M3	-0.0140	0.0002	B1T2M1	-0.0100	0.0001
	(0.0048)	(0.0000)		(0.0022)	(0.0000)
B1T1	-0.0176	0.0003	B1T1M2	-0.0048	0.0000
	(-0.0067)	(0.0000)		(0.0002)	(0.0000)
B1T2	-0.0099	0.0001	B1T2M2	-0.0006	0.0000
	(0.0008)	(0.0000)		(0.0030)	(0.0000)
B2T1	-0.0003	0.0000	B1T1M3	0.0001	0.0000
	(-0.0079)	(0.0001)		(0.0005)	(0.0000)
B2T2	-0.0056	0.0000	B1T2M3	-0.0005	0.0000
	(-0.0044)	(0.0000)		(-0.0001)	(0.0000)
B1M1	0.0372	0.0014	B2T1M1	0.0018	0.0000
	(-0.0010)	(0.0000)		(-0.0043)	(0.0000)
B1M2	0.0050	0.0000	B2T2M1	-0.0028	0.0000
	(0.0026)	(0.0000)		(-0.0050)	(0.0000)
B1M3	-0.0025	0.0000	B2T1M2	-0.0008	0.0000
	(-0.0006)	(0.0000)		(-0.0003)	(0.0000)
B2M1	0.0039	0.0000	B2T2M2	0.0015	0.0000
	(0.0028)	(0.0000)		(0.0008)	(0.0000)
B2M2	0.0012	0.0000	B2T1M3	0.0044	0.0000
	(0.0037)	(0.0000)		(-0.0005)	(0.0000)
B2M3	-0.0031	0.0000	B2T2M3	-0.0042	0.0000
	(0.0007	(0.0000)		(-0.0002)	(0.0000)
T1M1	-0.1535	0.0236			
	(0.0532	(0.0028)			

Note.
This table contains the Pearson product-moment correlation coefficient (*r*) between the dependent variable $LRMSE_3$ and each vector of orthogonal coefficients for ANOVA with planned comparisons. Also presented is the corresponding r^2. The effects listed include
B1 (linear trend associated with $\rho_{b_1 b_2}$), B2 (quadratic trend associated with $\rho_{b_1 b_2}$), T1 (linear trend associated with $\rho_{\theta_1 \theta_2}$), T2 (quadratic trend associated with $\rho_{\theta_1 \theta_2}$), M1 (Maximum likelihood vs. all Bayesian estimation methods), M2 (Bayesian sequential vs. other Bayesian estimation methods), M3 (Bayesian modal vs. Bayes EAP estimation methods), and their various interactions.
For the compensatory data, $SS_{effect} = 0.9680$, $SS_{error} = 0.0320$, $SS_{total} = 1.0000$.
For the noncompensatory data, $SS_{effect} = 0.9714$, $SS_{error} = 0.0286$, $SS_{total} = 1.0000$.

Figure 6

Trends Across $\rho_{\theta_1\theta_2}$ Levels as Reflected in *LRMSE* $_3$ Based on the Compensatory MIRT Model

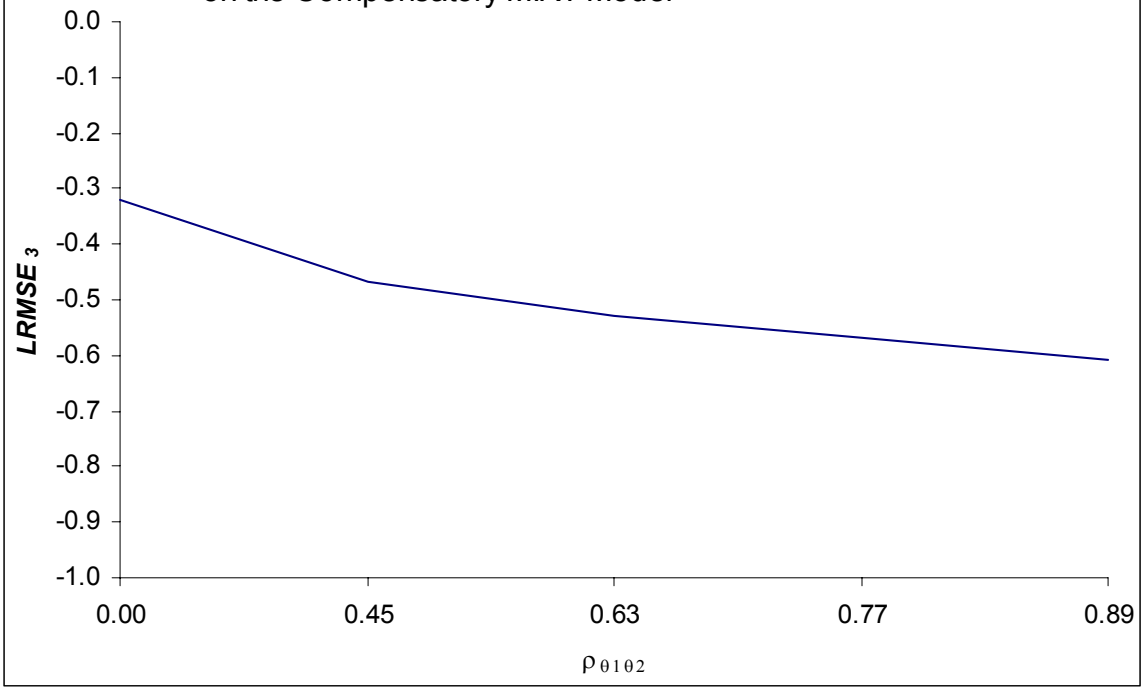


Figure 7

Differential Effect of METHOD on *LRMSE* $_3$ Based on the Compensatory MIRT Model

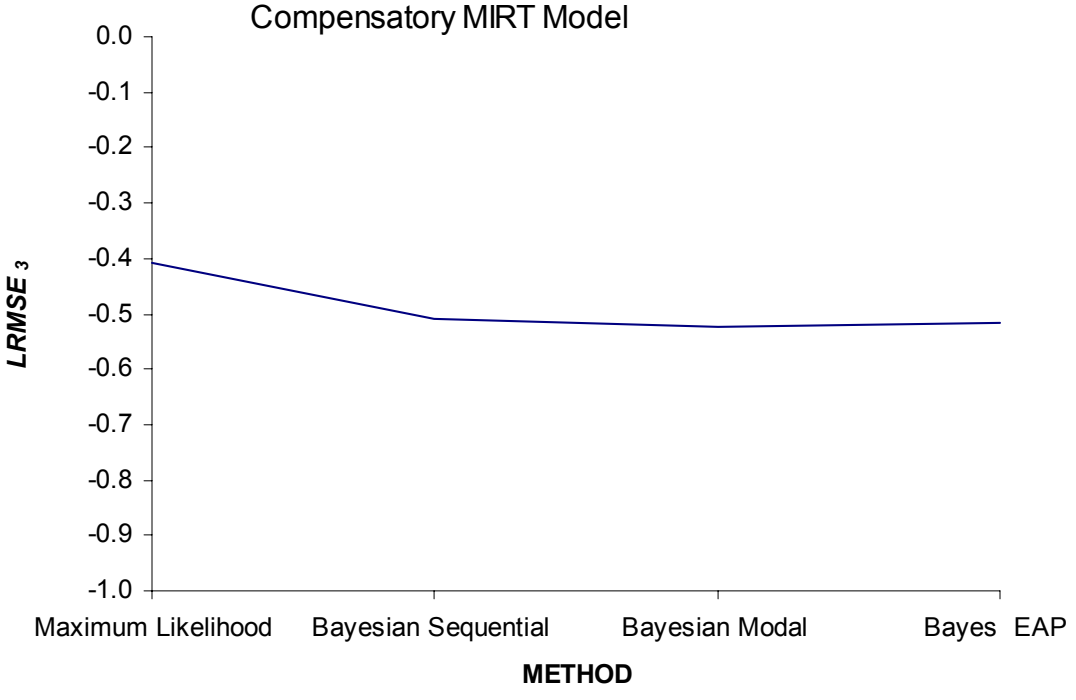


Figure 8

Trends Across $\rho_{\theta_1\theta_2}$ Levels as Reflected in *LRMSE* $_3$ Based on the Noncompensatory MIRT Model

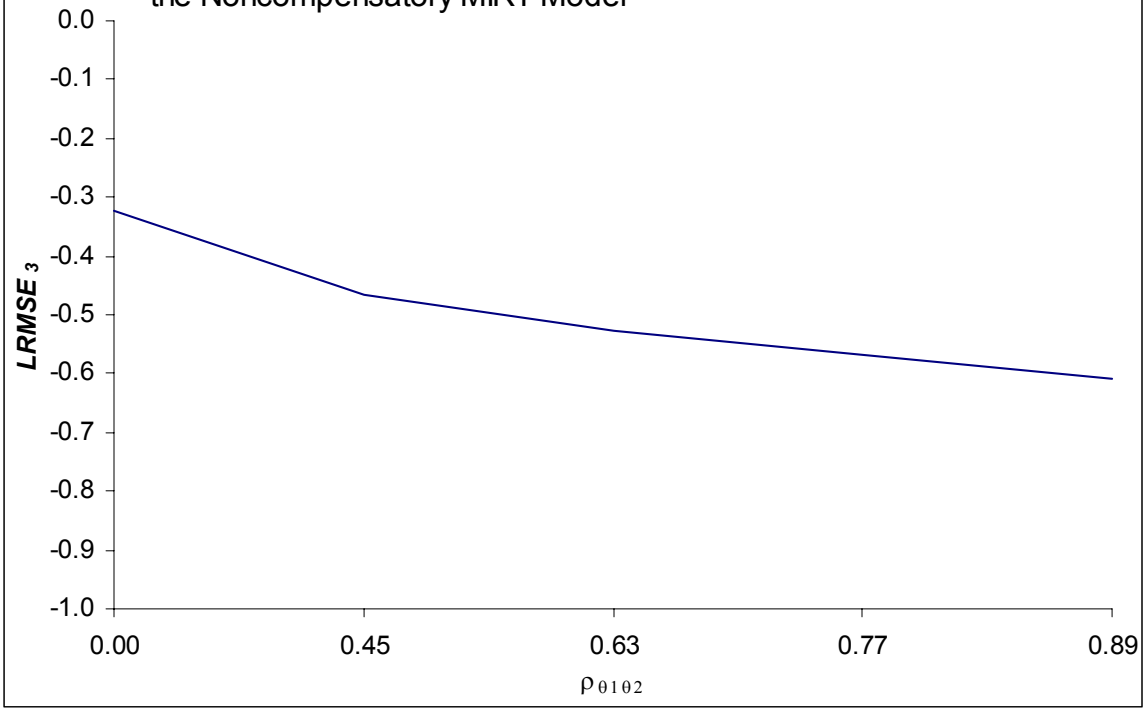


Figure 9

Differential Effect of METHOD on *LRMSE* $_3$ Based on the Noncompensatory MIRT Model

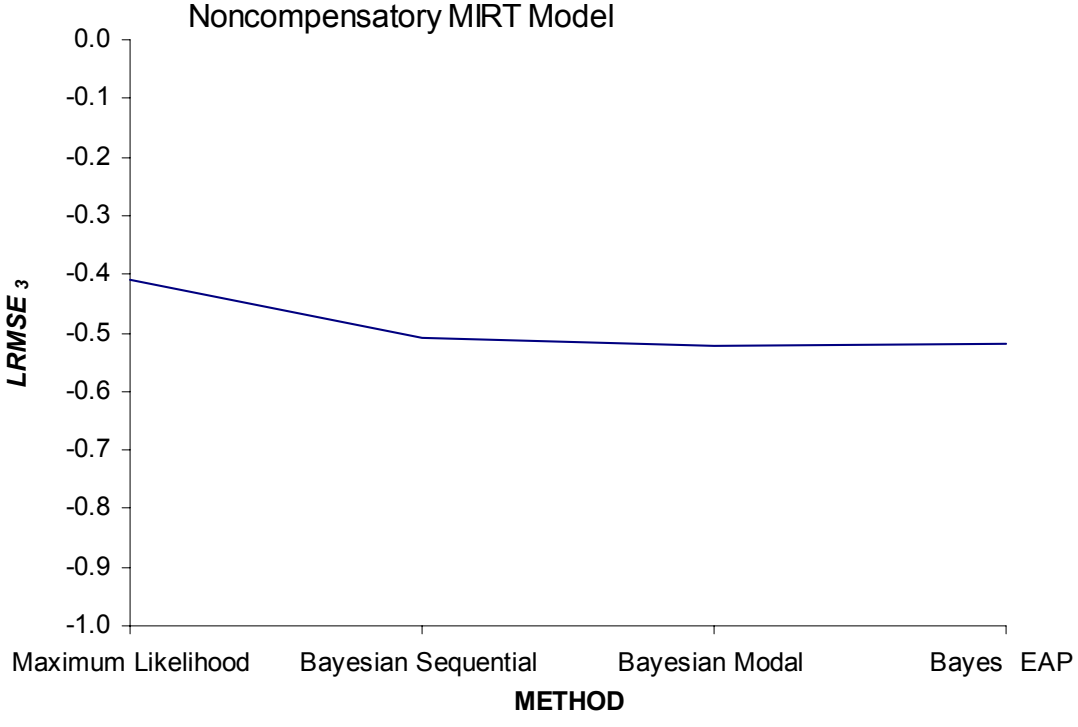


Table 9

Average Evaluative Indices of CAT Simulation Results by METHOD and Simulation Model

Item response matrix	Evaluative index					
	r	*rs*	*Bias*	*RMSE*	*ASE*	*ANI*
Maximum likelihood estimation						
Unidimensional	0.970	0.972	0.011	0.256	0.222	21.892
Two-dimensional						
Compensatory	0.944	0.949	0.016	0.390	0.227	21.002
Noncompensatory	0.908	0.918	0.022	0.515	0.302	28.267
Bayesian sequential estimation						
Unidimensional	0.971	0.974	0.005	0.238	0.211	21.444
Two-dimensional						
Compensatory	0.948	0.949	0.006	0.311	0.193	20.569
Noncompensatory	0.925	0.925	0.019	0.361	0.264	29.371
Bayesian Modal estimation						
Unidimensional	0.974	0.975	0.009	0.226	0.201	21.000
Two-dimensional						
Compensatory	0.952	0.951	0.003	0.300	0.182	20.667
Noncompensatory	0.924	0.923	0.042	0.364	0.258	27.466
Bayes EAP estimation						
Unidimensional	0.976	0.975	-0.005	0.220	0.211	21.487
Two-dimensional						
Compensatory	0.954	0.951	-0.008	0.304	0.194	20.927
Noncompensatory	0.927	0.926	0.010	0.363	0.269	28.514

Table 10

Average Evaluative Indices of CAT Simulation Results Based on the Unidimensional IRT Model and the Two-dimensional Compensatory MIRT Model by METHOD with $\rho_{\theta_1\theta_2}$ *Set At .89*

Item response matrix	Evaluative index					
	r	*rs*	*Bias*	*RMSE*	*ASE*	*ANI*
Maximum likelihood estimation						
Unidimensional	0.970	0.972	0.011	0.256	0.222	21.892
Two-dimensional	0.963	0.977	0.021	0.327	0.244	20.905
Bayesian sequential estimation						
Unidimensional	0.971	0.974	0.005	0.238	0.211	21.444
Two-dimensional	0.972	0.977	0.008	0.233	0.183	20.445
Bayesian Modal estimation						
Unidimensional	0.974	0.975	0.009	0.226	0.201	21.000
Two-dimensional	0.976	0.978	0.005	0.217	0.173	20.644
Bayes EAP estimation						
Unidimensional	0.976	0.975	-0.005	0.220	0.211	21.487
Two-dimensional	0.978	0.979	-0.007	0.213	0.186	20.875