Multidimensional Computerized Adaptive Testing in Recovering Reading and Mathematics Abilities

by

Yuan H. Li

Prince Georges County Public Schools, Maryland

William D. Schafer

University of Maryland at College Park

Address:
Yuan H. Li
Prince George's County Public Schools
Test Admin. Department, Room, 202E
Upper Marlboro, MD 20772
e-mail:yuanhwangli@juno.com

Paper presented at the annual meeting of the American Educational Research Association
April, 21-25, 2003, Chicago, IL.

Multidimensional Computerized Adaptive Testing in Recovering Reading and Mathematics Abilities

Abstract: Under a MIRT CAT's (multidimensional computerized adaptive testing) testing scenario, an ability estimate in one dimension will provide clues for subsequently seeking a solution in other dimensions. This feature may enhance the efficiency of MIRT CAT's item selection as well as its scoring algorithms compared with its counterpart, unidimensional CAT (UCAT). The shadow test approach (van der Linden & Reese, 1998) was imposed on both MIRT CAT and UCAT to maintain content balance for the current research design which intended to assess how feasibly (or well) the MIRT CAT can be implemented in a real testing program under current circumstances and to address the issue of which of these two CAT algorithms would perform better.

The present study used existing Reading and Math test data to generate simulated item parameters. A confirmatory item factor-analysis model was applied to the data using NOHARM (Fraser & McDonald, 1988) to produce interpretable MIRT item parameters. The item pool resulting from this analysis has two special features: (a) each Reading or Math item contributed only to its intended measure, and (b) since the simulated item discrimination parameters were intentionally defined for the Reading and Math dimensions, the MIRT CAT ability estimates should correspond to both measures, respectively, rather than to the abstract multidimensional abilities, $\underline{\theta}$.

Results showed that MIRT CAT conditional on the constraints was quite capable of producing accurate estimates on both measures. Compared with UCAT, MIRT CAT increased the accuracy of both ability estimates, especially for the low or high abilities in both measures, and reduced the rate of unused items in the item pool.

Key Words:

Computerized Adaptive Testing (CAT), Item Response Theory (IRT), Dimensionality, Zero-One Liner Programming, Constraints, Item Exposure, Reading Assessment, Mathematics Assessment

I. Introduction

A. Background of MIRT CAT

When a group of examinees takes a set of test items, test-examinee interaction might result in test data that appear to be unidimensional in some instances but multidimensional in otherss (Ackerman, 1992). A presumed single trait dimension for a test dataset that is actually multidimensional might jeopardize the invariant feature of item response theory (IRT) models (Ackerman, 1994; Reckase, 1985). Results from Li and Lissitz (2000) suggested that multidimensional IRT (MIRT) models can be applied to not only multidimensional data but also to unidimensional test data as well. The fit of MIRT models to unidimensional data will result in item discrimination estimates (or factor loadings) that approach zero for the overidentified dimensions but does little harm in terms of the IRT invariant feature. It seems apparent that MIRT models are more flexible for fitting test data than unidimensional models.

When MIRT is accommodated within the context of computerized adaptive testing (CAT), the result (MIRT CAT) enhances the efficiency of adaptive item selection as well as scoring algorithms (Luecht, 1996, Segall, 1996, 2000). The primary reason for such promising features has been documented by Segall (1996; 2000), who pointed out that "When the dimensions measured by a test or battery are correlated, responses to items measuring one dimension provide clues about the examinee's standing along other dimensions" (Segall, 2000, p53). Such a unique characteristic, that can not be fulfilled in the conventional unidimensional CAT (UCAT), might make MIRT CAT more appealing, such as by increasing the reliability for an examinee's ability estimate(e.g., Luecht, 1996, Segall, 1996).

With advances in the computing power of personal computers, CAT has gradually achieved popularity. For instance, UCAT has been implemented in large-scale testing programs such as the Graduate Record Examinations (GRE, 1993, Educational Testing Service). The process of extending CAT to multidimensional testing methods has also been explored (e.g., Boxom & Vale, 1987, Luecht, 1996, Segall, 1996, Tam, 1992, van der Linden, 1999, Veldkamp & van der Linden , 2002).

Several methods of MIRT CAT ability estimation were used by Bloxom and Vale (1987), Tam (1992), Luecht (1996), Segall (1996) and van der Linden (1999). Bloxom and Vale (1987) developed an approximate-scoring method with a closed form expression for point estimates of ability based on the multivariate extension of Owen's (1975) sequential updating procedure. Tam (1992) developed an iterative maximum likelihood (ML) method for the MIRT CAT ability estimation. Segall (1996), on the other hand, applied a Bayesian approach to the MIRT CAT ability estimation process by incorporating the prior distribution of multidimensional abilities.

Regarding the methods of MIRT CAT item selection by means of statistical criteria, several promising methods have been proposed (e.g., maximizing the determinant of the posterior information, [Segall, 1996], maximizing posterior expected kullback-leibler information, [Veldkamp & van der Linden , 2002]). When the estimation of a target composite score, as generated by a linear function of multidimensional abilities, is of interest, van der Linden (1999) introduced a minimum error variance criterion for item selection for that purpose.

Both ability estimation and item selection were simultaneously addressed by Segall (1996) and Luecht's (1996) studies. For the Segall study, Bayesian ability estimation and item selection by maximizing the DPI were employed. Within a licensing/certification program, Luecht (1996) extended Segall's approach by imposing a more complex set of content-balancing constraints. Both studies used the simple structure (i.e. where each item contributes to only one dimension and has zero loading on the remaining dimensions) to simulate MIRT item

discrimination parameters. In addition, the simulated MIRT CAT item parameters were adapted from the UCAT item parameters in order to compare UCAT and MIRT CAT.

Results from both studies indicated that a shorter MIRT CAT (about 25 % to 40%) could achieve about the same subscore reliability as its longer UCAT counterpart when multidimensional abilities are intercorrelated. Indeed, cross-information of an examinee's knowledge among various dimensions provides a better mechanism for choosing adaptive items for the examinee, whereas in multiple UCATs, cross-information among content areas is not utilized. If each of the simulated test items used in both studies has loadings on more than one dimension, more promising results in MIRT CAT might occur. However, as pointed out by a reviewer, the condition of an item loading on more than one dimension will complicate the interpretation of multiple-trait scores., Thus, a simple structure, where each item loads on a single dimension, seems more feasible especially when reporting multiple-trait scores rather than a composite score is of interest.

More sophisticated item-selection methods have not been extensively studied in the context of MIRT CAT. Van der Linden and Reese (1998) and van der Linden (1998, 2000) reviewed several algorithms for optimal assembly of tests from item banks. The utilization of a shadow test, as illustrated in van der Linden (2000), to be described later, represented one of the more promising methods for UCAT item selection. This method begins with assembling an on-line full test that meets all the desirable conditions (e.g., test specifications, constraints on dependencies between items) and has maximum information at a provisional ability estimate. An item with maximum information is then selected from the on-line shadow test, instead of directly from the item pool. Overall, advantages of using a shadow-test approach in CAT (van der Linden, 2000) are: (a) it has an optimal value for the objective function , (b) it is feasible (it can meet all required conditions without constraint violations; other methods might violate some of the constraint conditions), and (c) it is flexible (it can accommodate other, new demands, such as, Sympson and Hetter's [1985] item exposure control). An example of utilizing this item-selection method in MIRT CAT can be found in Veldkamp and van der Linden (2002).

B. Practical Problems of MIRT's Data Modeling

Although MIRT or its application in MIRT CAT has promising features over unidimensional IRT from a statistical viewpoint, MIRT has not yet been implemented in real testing programs. One primary reason is that unique solutions for estimated item loadings cannot be attained. The loadings are the coordinates on the axes (or dimensions) that define the space. Hence, rotating the axes will result in a new set of loadings. In addition, the interpretation of each dimension together with the determination of the number of dimensions for a given test dataset could be quite subjective, as is the case in factor analysis. The current MIRT computer software, TESTFACT 4 (Wood, Wilson, Gibbons, Schilling, Muraki & Bock, 2003), was programmed using the full information method. This program was designed to perform exploratory item-factor analysis although it might be used for a specific confirmatory factor analysis know as bi-factor analysis (Gibbons & Hedeker, 1992) which requires a factor pattern that should consist of one main factor plus group factors. This requirement might be limited to be used in some test data, but not suitable to all test data.

The confirmatory item factor-analysis modes in the MIRT context could help resolve these practical problems (McLeod, Swygert& Thissen, 2001). The FACT computer program (Segall, 1998), implementing the full information confirmatory item factor analysis using Markov chain Monte Carlo estimation, might be a suitable method for calibrating interpretable MIRT item parameters. The FACT program is, however, currently undergoing further investigation and is not available to the public. Existing available confirmatory item-factor analysis programs such as NOHARM (Fraser & McDonald, 1988) that models item-covariances rather than the full information approach has therefore become our choice for dealing with MIRT's item calibration. This program specifically allows constraints to be specified for the loading and/or covariances among the latent traits. Introducing the pattern matrices for those parameter matrices (e.g., indicating the values to be fixed at zero or at some other value) seems a flexible approach to seeking interpretable MIRT item parameters.

C. Research Purpose

As illustrated, UCAT might produce biased ability estimates when the condition of an under-fitted model occurs. Further, making use of ability estimates from other dimensions in item selection as well as in the scoring algorithms is precluded, making UCAT less efficient than MIRT CAT. With the compelling advantages of MIRT CAT (Segall, 2000), the intention in this study was to modify the current CAT to facilitate the process of locating Reading and Mathematics true abilities. The reason for choosing these two content areas as the example subject areas to be explored is that they are essential skills for students to be successful in all academic fields and they are often required to be taken at the same period of time in most testing programs. Of course, the methodology employed in these two content measures can be generalized to other content combinations (e.g., Reading and Science). Improving the accuracy as well as efficiency in assessing these two key content areas via MIRT CAT is the ultimate goal of this study.

A reviewer raised some key questions that most practitioners would want to address before committing to the additional complexities of MIRT CAT, or for that matter, before committing to UCAT over traditional liner testing. How much better does MIRT CAT perform than UCAT with comparable content constraints? How much better does MIRT CAT perform than conventional linear testing? Part of the latter issue has been addressed in Segall's (2001) study in which MIRT CAT was purposely designed to estimate a unique general ability. This provides a comparison between the magnitudes of score information resulting from MIRT CAT and number-right scoring methods (see Lord, 1980, p.73). However, if MIRT CAT was implemented for simultaneously estimating multiple abilities, a comparison between these two scoring methods may not exist. Accordingly, only the first issue (comparing MIRT CAT with UCAT) was studied in this research.

Due to the advantages of the shadow-test approach in CAT, it was adapted for use with with Segall's (2000) MIRT CAT item selection method. It is anticipated that validity of a test together with accurate ability estimates can be enhanced using this combination. Results from this empirical study can provide test practitioners a sense as to how feasible it is to implement MIRT CAT in a real testing program. This question was addressed by examining the accuracy of the multidimensional abilities as estimated by MIRT CAT under the testing situations described in the Methodology section.

II. Overview of Statistical Techniques

A. Multidimensional Logistic IRT Models

The model illustrated below is a multidimensional extension of the three-parameter logistic model (M3PL). This model hypothesizes that the probability of a correct response, $u_{ij}=1$, by person j to item i, given an individual's m-dimensional latent abilities, $\underline{\theta}_j$, is (refer to Reckase, 1985):

$$P(u_{ij}=1 \mid \underline{a}_i, d_i, c_i, \underline{\theta}_j) = c_i + (1-c_i)\frac{e^{Z_{ij}}}{1+e^{Z_{ij}}} \quad (1)$$

where,

$$Z_{ij} = D\left(\underline{a}_i' \underline{\theta}_j + d_i\right) \text{ or} \quad (2.1)$$

$$Z_{ij} = D\,\underline{a}_i'\left(\underline{\theta}_j - b_i \underline{1}\right) \quad (2.2)$$

D is a scaling constant (1.702),
$\underline{a}_i$ is a m-dimensional vector of item discrimination parameters,
c_i is a pseudo-guessing parameter,
d_i is an intercept parameter,
b_i is an item difficulty parameter, and
$\underline{1}$ is a m x 1 vector of 1's.

When Equation 2.1 is inserted into Equation 1, this becomes a commonly used form of the multidimensional model. This model can be reparameterized as the unidimenstional-like three-parameter model when Equation 2.2 is inserted into Equation 1 (Segall, 2001). Parameters d and b in Equations 2.1 and 2.2 are interchangeable, where $d = -b\sum_{i=1}^{m} a_i$ and $b = -d\left(\sum_{i=1}^{m} a_i\right)^{-1}$.

The location parameter of d is like an item difficulty which is a threshold parameter, given in the choice of model parameterization shown in Equation 2.1. The more common "difficulty" analog is computed as $\frac{-d}{\sqrt{\underline{a}_i' \underline{a}_i}}$ (Reckase, 1997). There is only a single location parameter rather than separate ones for each dimension. It would be preferable for each item to possess its own difficulty parameter for each dimension; however, multiple difficulty parameters for an item are mathematically indeterminate (Segall, 1996).

The scaling factor D is included in the model to make the logistic function as close as possible to the normal ogive function (Baker, 1992). Since the terms in Equation 2.1 or 2.2 are additive, being low on one latent trait can be compensated for by being high on the other latent traits. This model is, thus, called a compensatory model (Reckase, 1985). A multidimensional extension of the two-parameter logistic model (M2PL) is achieved if the guessing parameter c_i is constrained to zero for all items in Equation 1 above.

B. MIRT CAT Ability Estimates and Item Selection

1. The Likelihood Function

Assuming that the local independence assumption holds, given an examinee with multidimensional abilities $\underline{\theta}$ who responds to a set of n items with the response pattern $\underline{u}$, then the probability (or called likelihood function) of obtaining this response pattern $\underline{u}$ can be modeled by:

$$L(\underline{u} \mid \underline{\theta}) = \prod_{i=1}^{n} P(\underline{\theta})_i^{u_i} Q(\underline{\theta})_i^{1-u_i} \quad , i \in S_n \tag{3}$$

where $Q(\underline{\theta})_i = 1 - P(\underline{\theta})_i$ and S_n is the n items that have been administered (i.e., selected) to the examinee during the MIRT CAT testing process.

2. Posterior Density

The $\underline{\theta}$ parameter vector needs to be estimated. If the prior information $f(\underline{\theta})$ for the distribution (or probability density) of $\underline{\theta}$ together with the observed response pattern $\underline{u}$, are available, we are then able to approximate the posterior distribution of $\underline{\theta}$ according to the Bayes rule. The posterior density of $\underline{\theta}$ is:

$$f(\underline{\theta} \mid \underline{u}) = \frac{L(\underline{u} \mid \underline{\theta}) f(\underline{\theta})}{f(\underline{u})} \tag{4}$$

Where $f(\underline{u})$ is the marginal probability of $\underline{u}$ given by Bock and Lieberman (1970) and Bock and Aiken (1981):

$$f(\underline{u}) = \int_{-\infty}^{\infty} L(\underline{u} \mid \underline{\theta}) f(\underline{\theta}) d\underline{\theta} \tag{5}$$

$f(\underline{u})$ is irrelevant while finding the solution of the $\underline{\theta}$ parameter. Hence, the posterior function is proportional to a prior function times a likelihood function. That is:

$$f(\underline{\theta} \mid \underline{u}) \propto L(\underline{u} \mid \underline{\theta}) f(\underline{\theta}) \tag{6}$$

The relative influence of observed data (the input for the likelihood function) and prior information on the posterior function (related to the updated belief) depends on test-lengths, item-pool characteristics, and the magnitude of prior dispersion. As the prior becomes vague or diffuse, the posterior function is closely approximated by the likelihood function and consequently the Bayesian approach will result in the same solution as the likelihood approach. In contrast, if the prior is very informative or specific, then the prior distribution would have a relatively greater influence on the posterior function. Hence, consider relating this to the adaptive sequence which typically increases the peaked-ness of the likelihood function, reducing the influence of the prior.

3. Multidimensional Ability Estimates

As indicated previously, several methods of MIRT CAT ability estimation exist. Segall's (1996, 2000) model estimation of $\underline{\theta}$, denoted by $\underline{\hat{\theta}}$, is briefly introduced here. The estimates are those values that maximize the posterior density function of $f(\underline{\theta}|\underline{u})$. The $\hat{\theta}$ can be derived by partially differentiating the log-posterior density function with respect to the m dimensional $\underline{\theta}$, setting these equal to a zero vector (Equation 7), and simultaneously solving these m non-linear equations using the Newton-Raphson method or the Fisher method of scoring (refer to Segall, 1996, 2000).

$$\frac{\partial}{\partial \underline{\theta}} \ln f(\underline{\theta}|\underline{u}) = \underline{0} \tag{7}$$

For example, in the two-dimensional case, the nonlinear Equation 7 can be resolved for θ_1 and θ_2 iteratively until the differences between two successive solutions (t and t+1) become sufficiently small. The iterative equation is given below.

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix}_{t+1} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix}_{t} - \left[-\mathbf{I}\left(\underline{\hat{\theta}}\right) \right]^{-1} \cdot \begin{bmatrix} \frac{\partial \ln}{\partial \theta_1} \ln f(\underline{\theta}|\underline{u}) \\ \frac{\partial f}{\partial \theta_2} \ln f(\underline{\theta}|\underline{u}) \end{bmatrix}_{t} \tag{8}$$

where the negative posterior information matrix, $-\mathbf{I}(\underline{\theta})$, is computed from equation 9 (below).

The Fisher method was chosen for the present study to ensure convergence within the iterative steps of ability estimation. The Newton-Raphson method might not converge if the initial starting values of $\underline{\theta}$ are far removed from their final-stage estimates.

4. MIRT CAT Item Selection

When the Bayesian approach is used for estimating a unidimensional ability estimate, the items that contribute maximum information to the posterior are selected, allowing the error variance to be reduced with fewer items. Similarly, the criterion for MIRT CAT item selection proposed by Segall (1996, 2000) is to seek an item with the maximum determinant of the posterior information matrix (DPI) value. An item with the maximum DPI will result in "the largest decrement in the volume of the credibility ellipsoid" (the Bayesian measure of uncertainty, Segall, 1996, p.341). This statistical concept has been graphically illustrated in Segall's article (2000). The DPI consists of three elements (refer to Segall, 2000):

$$I_{i|S_{n-1}} = \Phi^{-1} + W_i + W_{S_{n-1}} \quad (9)$$

Where the subscript i represents the candidate item, and
S_{n-1} is the notation for the previously administered items.

The first element is the inverse prior covariance matrix of multidimensional abilities. The second element consists of the **W**-matrix for the candidate item i. The third element (see Equation 9) is the sum of **W**-matrices across those previously administered items.

$$W_{S_{n-1}} = \sum_{j \in S_{n-1}} W_j \, . \quad (10)$$

All the **W**- matrices, including the candidate and the previously administered items, are defined by the following equation (Segall, 2000):

$$W_i = D^2 \, a \, a' \, w_i^* \quad (11)$$

Where

$$w_i^* = \frac{Q_i(\theta)}{P_i(\theta)} x \left[\frac{P_i(\theta) - c_i}{1 - c_i} \right]^2 \quad (12)$$

Overall, each candidate-item's DPI is the determinant value of the matrix that is defined by Equation 9.

Instead of maximizing the DPI, Veldkamp and van der Linden (2002) proposed another item-selection method by maximizing posterior expected Kullback-Leibler (KL) information. Their method is an application of extending KL information on UCAT (proposed by Chang and Ying, 1996) to MIRT CAT. Conceptually, an item with maximum KL information will serve the same purpose as maximum DPI, even though both criteria were derived from different mathematical functions. One advantage of this KL-based method is that, in theory, it takes into consideration the uncertainty of ability estimates in the process of item selection so that it will improve the efficiency of item selection, especially in early stages of CAT (Chang & Ying, 1996). However, the true $\underline{\theta}$ parameter vector that is part of the KL information function is unknown, so the computation of the KL information, as well as DPI, depends on the current estimated $\underline{\theta}$ parameter vector, which is poorly estimated at the early-stage MIRT CAT. This factor will have a negative effect on the performance of KL-based item selection method, as well as the Fisher information method (e.g. DPI). Based on the study conducted by Veldkamp and van der Linden (2002), the differences in accuracy of multidimensional ability estimates resulting form both item selection methods seems to be negligibly small. In addition, the overlap of items selected by both methods was about 90 percent. Their work also illustrated how to apply the KL-based item selection method to different cases of multidimensionality (e.g., all abilities intentional, intentional and nuisance abilities, etc.)

C. Constrained MIRT CAT with Shadow Tests

1. Zero-One Liner Programming

Each item's DPI value is the function of an examinee's provisional $\underline{\hat{\theta}}$ and its value will be updated along with the examinee's updated $\underline{\hat{\theta}}$ during the MIRT CAT process. If seeking the most accurate ability estimates is the only factor to be considered, an item with the maximum DPI is one of the most preferred criteria to be used for MIRT CAT item selection. However, if other factors such as content balance and item exposure control are also needed to be taken into account, the technique of zero-one linear programming (Theunissen, 1985, 1986; van der Linden & Boekkooi-Timminga, 1989) is a suitable method that can be easily and effectively adapted into the MIRT CAT process. A description of zero-one linear programming is presented below, before we illustrate how to assemble shadow tests (van der Linden & Reese, 1998) that utilizes the zero-one linear programming.

Linear programming is designed to seek the maximum value for a linear function such as Equation 13 while the required constraints formalized in Equation 14 are imposed.

$$\text{maximize } \sum_{i=1}^{L} \mathrm{DPI}_i(\underline{\hat{\theta}})x_i, \tag{13}$$

subject to

$$\mathbf{A} \cdot \underline{x} \square \underline{b} \tag{14}$$

where $\square$ could be $<$, $=$, or $>$,

and for i to L,

$$x_i \in \{1,0\} \tag{15}$$

In Equation 14, the items in the bank are indexed by i=1,…,L and the values in the variable x_i are parameters that will be estimated. For zero-one linear programming, the x values are constrained to be either one or zero as indicated in Equation 15 to identify whether an item is a qualified candidate item. The value of one or zero in the decision variables, x, indicates whether the items are selected or not for the shadow test.

A vector will be created as shown below before seeking the solution---the binary decision values for the x-vector. The matrix of **A** together with vector of $\underline{b}$ are used to specify what constraints are specified. In the present example we have an item pool with 10 items; the first five items belong to the first domain and the rest of the items belongs to the second domain. In addition, there are two constraints to be imposed, namely 2 and 3 items from these two domains, respectively, are expected to be administered to each examinee. Under this testing scenario, the **A** matrix and $\underline{b}$ vector will be created as shown below before seeking the solution of the vector parameter $\underline{x}$.

$$\mathbf{A}=\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\underline{b} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

The number of columns in **A** matrix should equal the number of items in the pool. In addition, each row in **A** matrix together with the corresponding row in $\underline{b}$ vector expresses a single constraint. The first constraint as described above is expressed in the first row in **A** matrix together with the first row in $\underline{b}$ vector. The five series of "1" connote that the first five of the 10 items are Domain 1 and the last five series of "0" connote that the last five of the 10 items are not Domain 1. Finally, the condition of 2 items in Domain 1 to be picked is specified as "2" in the first row in $\underline{b}$ vector.

2. MIRT CAT Using Shadow-test Approach

The shadow-test CAT was proposed by van der Linden and Reese (1998) using the technique of the zero-one linear programming to impose all types of constraints. Various types of constraints that can be employed in CAT are enumerated in van der Linden and Ress (1998) and van der Linden (2000). The algorithms (refer to van der Linden & Reese, 1998; van der Linden, 2000) for the MIRT CAT in the context of shadow test are:

(1). Set the initial ability estimates (e.g., a vector, that can be randomly generated from a multivariate normal distribution, MNV($\underline{0}$, $\boldsymbol{\Phi}$), where $\boldsymbol{\Phi}$ is prior variance-covariance of $\underline{\theta}$).

(2). Assemble an on-line shadow test that:
(a) has met all the constraints as specified by Equation 14,
(b) maximizes DPI value on the Equation 13 at the provisional ability estimates, and
(c) includes the previously administered item(s).

(3). Administer the item with maximum DPI among items from the on-line shadow test.

(4). Re-estimate abilities based on the examinee's response(s) to the items that have been administered.

(5). Release all unused items that have been previously included in the shadow test to the item pool.

(6). Add an additional constraint to the constraints that have been imposed in the zero-one linear model to ensure that the item being recently selected and administered to the examinee "must" be included in the next updated on-line shadow test. For matrix expression in the above example, if Item 2 is selected first, the **A** matrix and $\underline{b}$ vector should be updated in the following way:

$$\mathbf{A}=\begin{bmatrix} 1&1&1&1&1&0&0&0&0&0 \\ 0&0&0&0&0&1&1&1&1&1 \\ 0&1&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$$\underline{b}=\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

The last row in A matrix combined with the last row in $\underline{b}$ vector indicates that Item 2 is certain to be part of the updated shadow test.

(7). Repeat the procedures 2-6 until the fixed n items (or other criteria) have been administered (or met).
Thus, the last shadow test is the actual adaptive test that meets all desirable constraints.

D. Calibrate MIRT CAT Interpretable Item Parameters

As noted earlier, the underlying dimensions of test data are sometimes hard to define. This makes it difficult to find the interpretable values of item loadings. A test battery of CTBS (CTB/McGraw-Hill, 1997) is designed to measure multiple content areas for second grade students, in which the Reading and Math Concept tests were selected in this study. These two content-area scores are presumed to constitute the underlying dimensions of the test data collected from students' responses on these two tests. This assumption would simplify the application of MIRT on CAT. These two content-area scores are required to be estimated and reported (e.g., for instructional purposes) during the process of MIRT CAT.

An item response matrix from 9351 examinees on the 25 Reading and 26 Math items on the CTBS test was used for NOHARM item calibration. For the NOHARM program settings, a Math Concept item is only allowed to load on its intended-measured Math dimension but not the secondary-minor Reading dimension. A Reading item, in contrast, is only allowed to load on the Reading dimension, not on the secondary-minor Math dimension. In addition, because the NOHARM has difficulty in estimating the guessing parameters, the guessing parameters were first estimated by BILOG (Mislevy & Bock, 1990) and then supplied as predetermined (or fixed parameters) to the NOHARM program. Finally, an oblique solution was used to maintain interpretability of correlated trait scores.

The results produced by NOHARM are presented in Table 1. The bottom of Table 1 also presents the summary statistics of the MIRT item parameters. The correlation between Reading and Math dimensions is .729 which was subsequently substituted into the prior covariance matrix Φ required for adaptive item selection and scoring. In terms of Data-model fit for this particular solution, the root mean square of residuals between the model-based and the observed covariance is .0053.

Table 1:
MIRT Item Parameters Produced by NOHARM and BILOG

Reading Items	d	a_1	a_2	c	Math Items	d	a_1	a_2	c
1	0.390	0.621	0	0.152	26	1.475	0	0.361	0.246
2	-1.061	1.190	0	0.275	27	0.851	0	0.515	0.129
3	0.294	0.778	0	0.168	28	-0.678	0	1.078	0.271
4	-0.760	1.627	0	0.450	29	0.396	0	0.809	0.418
5	1.533	1.056	0	0.207	30	1.545	0	0.433	0.308
6	0.873	1.411	0	0.450	31	0.381	0	1.069	0.371
7	0.878	0.482	0	0.450	32	0.845	0	0.818	0.429
8	1.174	0.963	0	0.224	33	-0.332	0	0.811	0.101
9	0.912	0.700	0	0.291	34	-0.293	0	0.786	0.249
10	1.333	0.568	0	0.271	35	0.813	0	0.709	0.237
11	0.475	0.751	0	0.158	36	0.648	0	0.802	0.218
12	0.552	0.895	0	0.127	37	0.695	0	0.837	0.375
13	0.726	1.044	0	0.041	38	1.399	0	0.947	0.155
14	0.233	0.804	0	0.211	39	-1.013	0	0.871	0.340
15	-1.114	0.545	0	0.283	40	-0.183	0	0.467	0.464
16	-1.841	1.933	0	0.428	41	0.690	0	0.889	0.135
17	-1.964	1.686	0	0.230	42	0.471	0	0.450	0.235
18	0.717	0.775	0	0.076	43	-0.196	0	0.764	0.218
19	0.280	0.790	0	0.326	44	0.259	0	0.992	0.221
20	-1.284	0.942	0	0.138	45	-1.630	0	0.835	0.137
21	0.530	1.563	0	0.472	46	-0.553	0	1.463	0.145
22	-2.176	1.386	0	0.223	47	0.536	0	1.212	0.391
23	0.970	1.209	0	0.150	48	0.739	0	0.719	0.361
24	0.360	0.687	0	0.253	49	-1.601	0	0.709	0.126
25	0.278	0.913	0	0.192	50	-1.088	0	0.805	0.209
N/A					51	-0.190	0	0.768	0.450
Mean	0.092	1.013	0.000	0.250	Mean	0.153	0.000	0.805	0.267
SD	1.072	0.393	0.000	0.122	SD	0.879	0.000	0.245	0.112
Minimum	-2.176	0.482	0.000	0.041	Minimum	-1.630	0.000	0.361	0.101
Maximum	1.533	1.933	0.000	0.472	Maximum	1.545	0.000	1.463	0.464

III. Methodology

A. Simulated Item Bank and Ability Parameters

Fifty-one sets of MIRT parameters were obtained using the procedures described above. However, 51 items are not enough to form an item pool. This test's MIRT item parameters are presumed population parameters. By introducing sampling error into this test's item parameters, we are able to generate as many tests' item parameters as we wanted. More specifically, for resampling another test's item parameters, the set of the item parameters for the i^{th} item is sampled from a (m+1)-dimensional normal distribution, MVN(μ_i, Σ), where μ_i =(d_i, ln a_{1i}, ln a_{2i})

and the variance-covariance matrix $\sum$ was estimated from the transformed estimates of the 51 items. Afterwards, the generating values that represent ln a_{1i} and ln a_{2i} were converted back to their original units using "exponential" transformation. The above procedure was repeated for each test item and repeated for 15 simulated tests.

Finally, we obtained a 765-item pool whose summary of item characteristics is presented in Table 2. Compared the summary statistics for item characteristics from the population parameters as presented in Table 1, the resulting 765 items are more diverse in terms of the variance of difficulty and discrimination parameters. Because the characteristic of the intercorrelation of item parameters for an item exists in real test data, we believe that this way of generating an item pool can retain this feature although the approach to generating an item pool relies more heavily on the distributional assumption. Of course, the simulated item parameters calibrated from a real item pool is the best choice if an appropriate MIRT item pool exists.

Table 2:
Summary of Item Characteristics for the Simulated Item Pool

Item Parameter	N	Mean	SD	Minimum	Maximum
Difficulty	765	.16	1.34	-4.62	4.01
Discrimination for Reading	375	1.07	.61	.22	3.82
Discrimination for Math	390	.85	.42	.23	3.70
Guessing	765	.26	.12	.04	.47

Five hundred simulees were randomly selected from MVN ($\underline{0}$, $\mathbf{\Phi}$), where $\mathbf{\Phi}$ is the population varaince-covarance matrix of the Reading and Math dimensions. The $\mathbf{\Phi}$ is obtained from the NOHARM. The covariance value of both content-area scores was .729 which is the off-diagonal value of $\mathbf{\Phi}$ and the both diagonal values of $\mathbf{\Phi}$ are 1. This sample was used for evaluating the accuracy of MIRT CAT's ability estimates. Similarly, another five thousand simulees were generated and then used for evaluating the exposure rate of the item pool items.

B. Ability Estimates and Item Selection

The Bayesian model approach (Segall, 1996) was chosen for estimating multidimensional abilities. This method will incorporate the prior information of highly correlated Reading and Math measures into the likelihood function so that an ability estimate (e.g., Reading) in one dimension will provide clues for subsequently seeking a solution in other dimensions (e.g., Math). The $\mathbf{\Phi}$ matrix, as mentioned above, was also used for the prior varaince-covaraince while estimating abilities. The initial abilities of all simulated subjects at the beginning of the test were taken as an vector that was randomly drawn from a multivariate normal distribution, MNV($\underline{0}$, $\mathbf{\Phi}$). The MIRT CAT stopped when the fixed test length was reached.

The maximum DPI criterion (Segall, 2000) together with the shadow-test approach (van der Linden, 2000) was chosen for item-selection method in this study because the maximum DPI criterion selects items to maximize accuracy along all dimensions simultaneously (Segall, 2001), and the shadow-test approach ensures that content balance was achieved.

C. Simulation Conditions

Two simulation conditions are listed in Table 3. There were 29 constraints imposed to assemble an on-line shadow test. These 29 constraints (listed in Appendix A) corresponded to the 29 objectives of the test specifications that were used for editing the CTBS Reading and Math tests. Hence, the number of items for each objective on the shadow test was constrained as in the original test. In addition, the 25 Reading items were administered to simulees, first, followed by 26 Math items.

The second simulation condition modeled the application of UCAT algorithms to the multidimensional item pool, in which each Reading or Math item has its own unidimensional-like item parameters (e.g., a, b, and c). Technically, the second simulation was conduced only if items were allowed to load on only one dimension (simple structure, as shown on Table 1) and the prior distribution of multidimensional abilities, $\mathbf{\Phi}$, was specifically set at a diagonal matrix (e.g., $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ used in this study, refer to Segall, 1996, p.347). When these two conditions held, the negative posterior information matrix in Equations 8 and 9 becomes a diagonal matrix. Inverting a diagonal matrix in Equation 8, involves only taking the reciprocals of the elements on the main diagonal and using that new diagonal matrix as the inverse. Afterwards, the matrix multiplication in Equation 8 is merely a diagonal. As a result, the value of a θ_m (e.g., θ_1) solution does not affect that for another θ_m (e.g., θ_2) solution, and vice versa. In other words, the solution to the m equations given by Equation 7 can be solved for each θ_m (for m=1, 2, …m) separately and each θ_m is the result of unidimensional Bayesian modal estimation.

Additionally, the DPI item selection method will result in the same solution as its counterpart of the univariate maximum likelihood item information function (Segall, 1966, p. 347).

D. Evaluation

One hundred replications for each condition were conducted. The BIAS along with the RMSE (root mean squared error) statistics of the ability estimates across these two simulation conditions were computed by the formulas shown below.

$$\text{BIAS}(\theta_i) = \frac{\sum_{i=1}^{r} (\hat{\theta}_i - \theta_i)}{r} \quad \text{and} \quad (16)$$

$$\text{RMSE}(\theta_i) = \sqrt{\frac{\sum_{i=1}^{r} (\hat{\theta}_i - \theta_i)^2}{r}} \quad (17)$$

where θ_i is the true ability parameter, $\hat{\theta}_i$ is the corresponding estimated ability parameter, and r is the number of replications, which was 100 in this study.

RMSE is a measure of total error of estimation that consists of the systematic error (BIAS) and random error (SE). These three indexes relate to each other as follows (Rao, 2000):

$$\text{RMSE}(\theta_j)^2 \cong \text{SE}(\theta_j)^2 + \text{BIAS}(\theta_j)^2 \quad (18)$$

As can be seen from Equation 18, either a large variance (SE^2) or a large BIAS will produce a large RMSE. It is apparent that an estimator will have much practical utility only if it must not only be highly precise (or small SE^2), but also has small BIAS (Rao, 2000). The accuracy of an estimator is inversely proportional to its RMSE so that this RMSE index is the criterion of accuracy for an estimator (Rao, 2000). Accordingly, this index was primarily used to compare the accuracy of ability estimates when they were estimated under various simulation conditions.

The item-exposure rate refers to the ratio of the number of times an item has been administered to the total number of test-takers. The following indices (refer to Revuelta & Ponsoda, 1998) will be used to compare the five CAT algorithms: (a) the percentage of items never administered in the population, (b) the standard deviation (SD) of the variable of the item-exposure rate, (c) the minimum and maximum values of this variable. The distribution of the item-exposure rates, grouped in several intervals, will be also computed for each CAT condition.

Table 3. Research Conditions being Simulated

Independent Variable	Options		Dependent Variables
Item Selection	MIRT CAT: Assembling an on-line shadow test by maximizing the DPI	UCAT to the multidimensional item pool	Ability Estimates Item Exposure Rates
Test Length	25 Reading and 26 Math Items	25 Reading and 26 Math Items	

E. Computer Program

The computer program MIRTCAT was used for running the simulation conditions. The MIRTCAT was coded by the MATLAB matrix language (The MathWorks, 2001), in which the 0-1 linear programming was resolved from the callable library of LINDO API (LINDO Systems, Inc. 2001). Technically, the LINDO API was called into the MIRTCAT to seek the solution of the vector of $\underline{x}$ in Equation 12. All solutions were met all the constraints without any difficulties, as well as in a timely manner (e.g. less a second per item).

IV. Results and Discussions

A. Multidimensional Ability Recovery under MIRT CAT

1.Comparison Between the MIRT CAT and UCAT Conditions

The first research condition applied the shadow test approach on MIRT CAT by maximizing the DPI. The second research condition was formed by using the UCAT algorithms on the same multidimensional item pool.

Table 4 shows mean, standard deviation (SD), minimum and maximum of BIAS and RMSE in the recovery of Reading and Math abilities under the two conditions. The summary statistics presented in Table 4 seem to show that MIRT CAT tends to result in more accurate scores than UCAT did. For instance, the average RMSEs for Reading were .212 and .219 for the MIRT CAT and UCAT; the average RMSEs for Math were .276 and .301 for the MIRT CAT and UCAT. The minimum RMSEs of Reading were .121 and .135 for the MIRT CAT and UCAT conditions, respectively; the minimum RMSEs of Math were .170 and .173 for the MIRT

CAT and UCAT conditions. In terms of maximum RMSEs of Reading, they were .515 and .634 for the MIRT CAT and UCAT. A large difference in the maximum RMSEs of Math between MIRT CAT and UCAT was found; they were .891 and 1.197, respectively.

Table 4: Descriptive Statistics of Bias and RMSE indices of Multidimensional Ability Estimates under the shadow-test constraint and the shadow-test together with UCAT (N=500).

Condition	Ability	BIAS				RMSE			
		Mean	SD	Minimum	Maximum	M	SD	Minimum	Maximum
MIRT CAT	Reading	.031	.085	-.435	.368	.212	.059	.121	.515
	Math	.034	.143	-.648	.819	.276	.089	.170	.891
UCAT	Reading	.027	.084	-.601	.441	.219	.066	.135	.634
	Math	.040	.156	-.595	1.087	.301	.120	.173	1.197

Figure 1a presents a plot of BIAS in Reading as a function of true Reading abilities for the two research conditions. Figure 1b shows the difference of the absolute value of UCAT BIAS minus from the absolute value of MIRT CAT BIAS. A positive value in Figure 1b implies that UCAT produced more absolute value of BIAS, and vice versa. Figure 1a or 1b showed that MIRT CAT tended to produce less BIAS for the low and high reading abilities. Figure 1b showed that UCAT tended to produce a smaller absolute value of BIAS in many spots of middle range of reading abilities.

Figure 1c presents a plot of RMSE in Reading as a function of true Reading abilities for the two research conditions. Figure 1c showed that MIRT CAT tended to produce less RMSE for the low and high reading abilities. This can also be seen in Figure 1d, which shows the difference of the value of UCAT RMSE minus from the value of MIRT CAT RMSE. A positive value in Figure 1d implies that UCAT produced more RMSE value, and vice versa. Figure 1d showed that both UCAT and MIRT CAT produced almost the same amount of RMSE in the middle range of reading abilities.

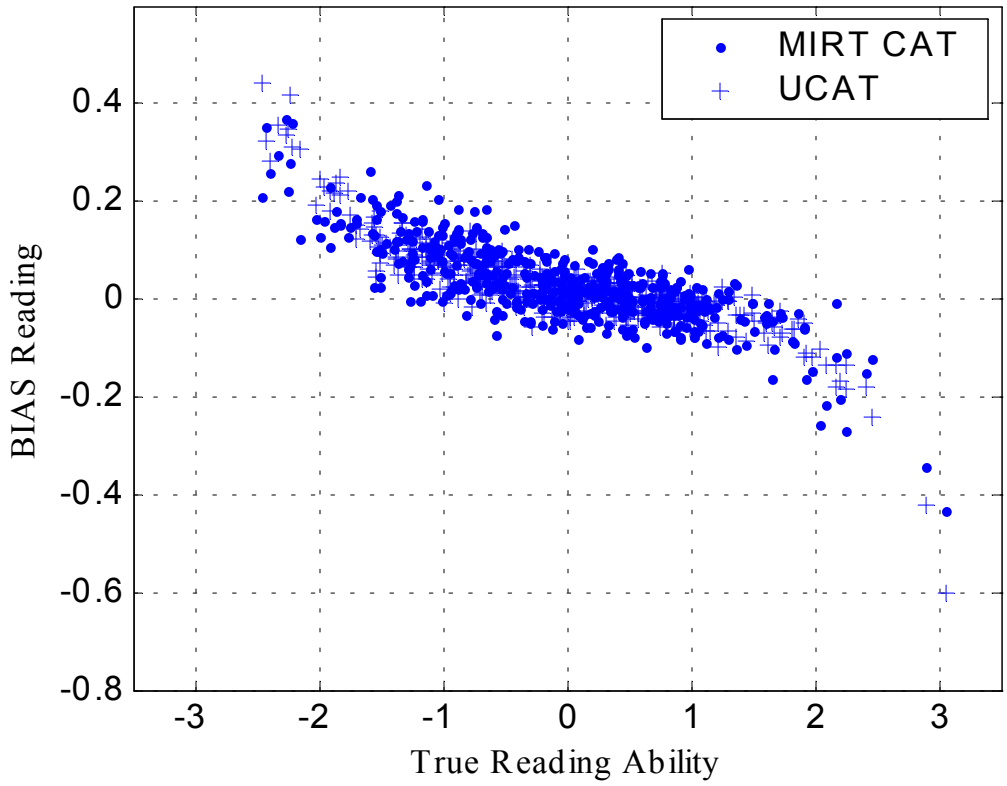

Figure 1a BIAS as a Function of True Reading Ability for MIRT CAT and UCAT

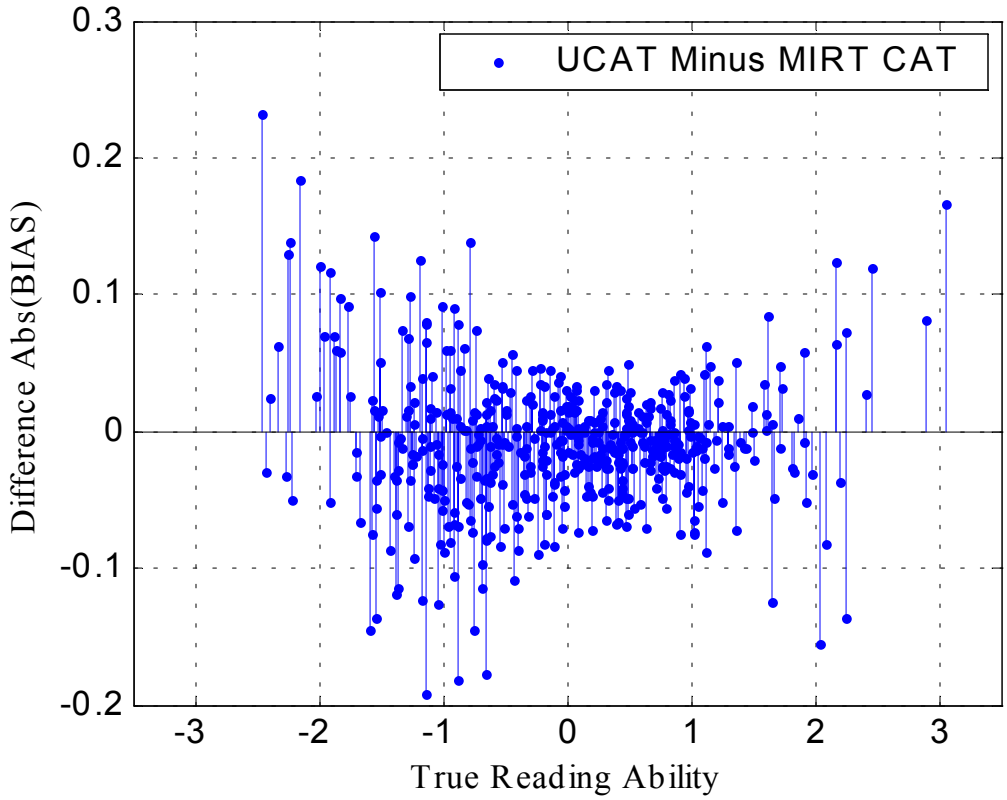

Figure 1b Difference Abs(BIAS) as a Function of True Reading Ability

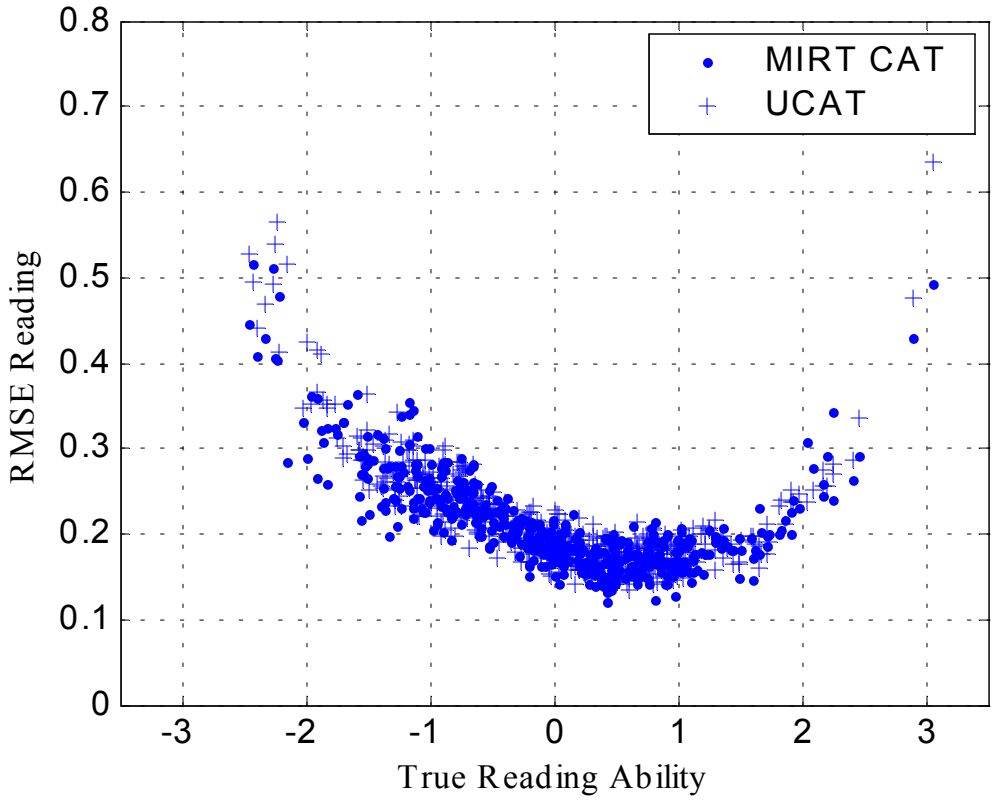

Figure 1c RMSE as a Function of True Reading Ability for MIRT CAT and UCAT

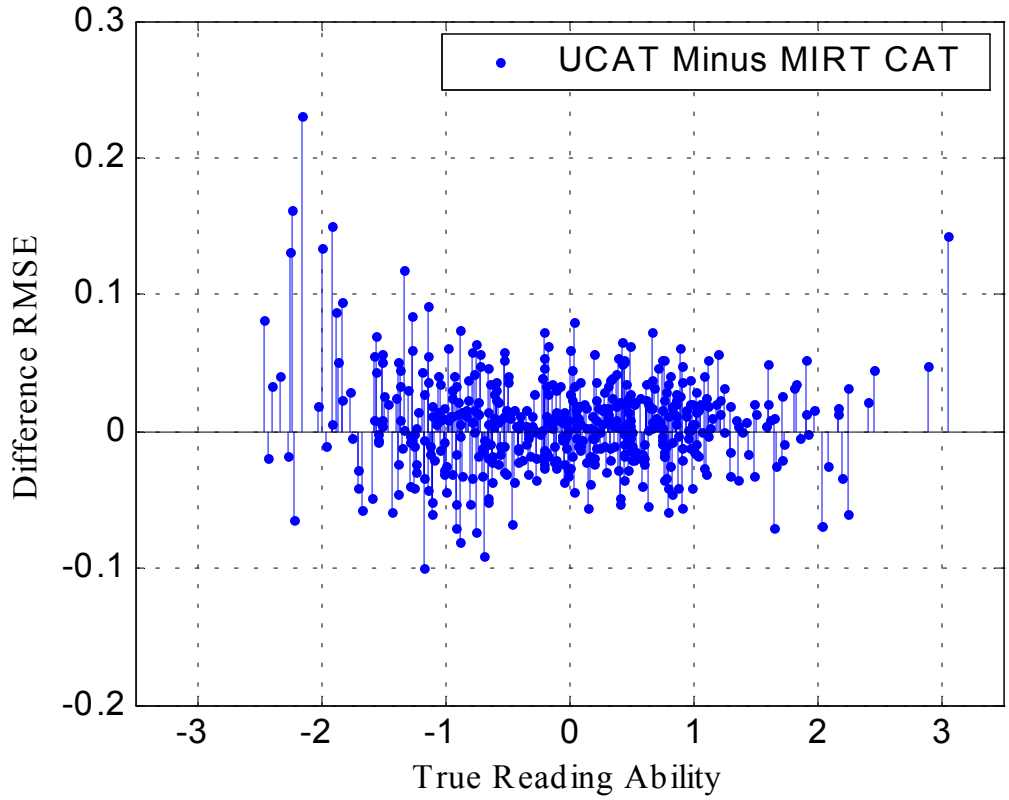

Figure 1d Difference RMSE as a Function of True Reading Ability

Figure 2a presents a plot of BIAS in Math as a function of true Math abilities for the two research conditions. Figure 1a showed that MIRT CAT tended to produce less BIAS for the low and high Math abilities. This can also be seen in Figure 1b, which shows the difference of the absolute value of UCAT BIAS minus from the absolute value of MIRT CAT BIAS. Figure 2b shows that UCAT tended to produce less absolute BIAS in many regions of middle range Math abilities.

Figure 2c presents a plot of RMSE in Math as a function of true Math abilities for the two research conditions and Figure 2d shows the difference of the value of UCAT RMSE minus from the value of MIRT CAT RMSE. Figure 2c or 2d shows that MIRT CAT tended to produce less RMSE.

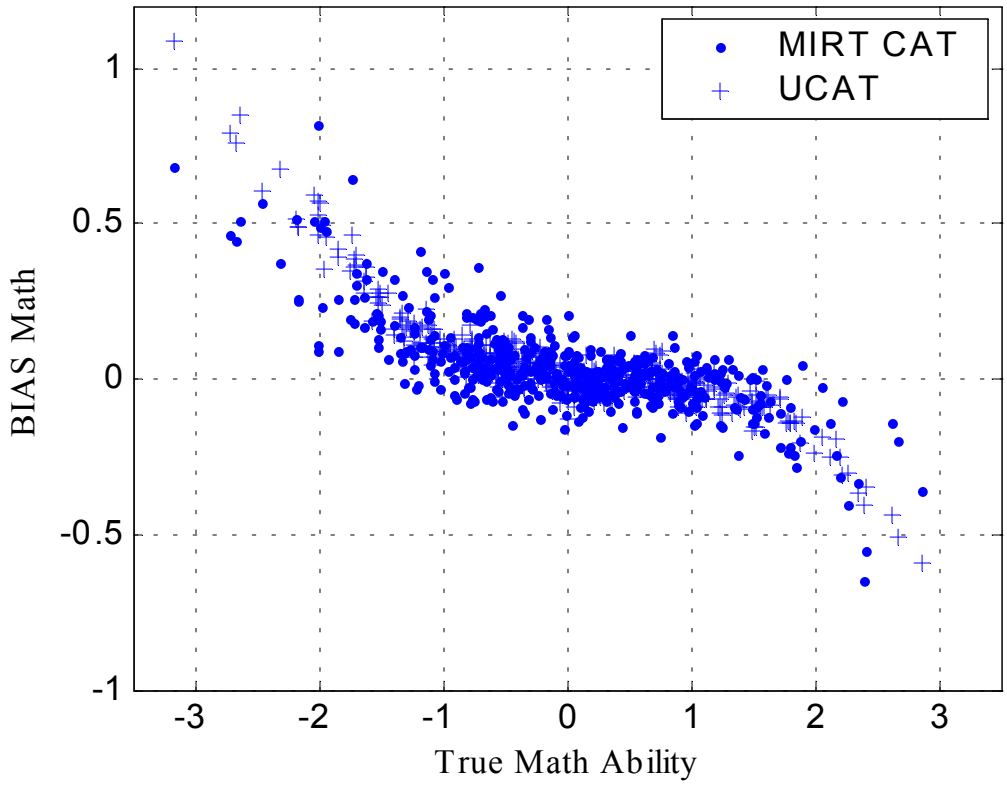

Figure 2a BIAS as a Function of True Math Ability for MIRT CAT and UCAT

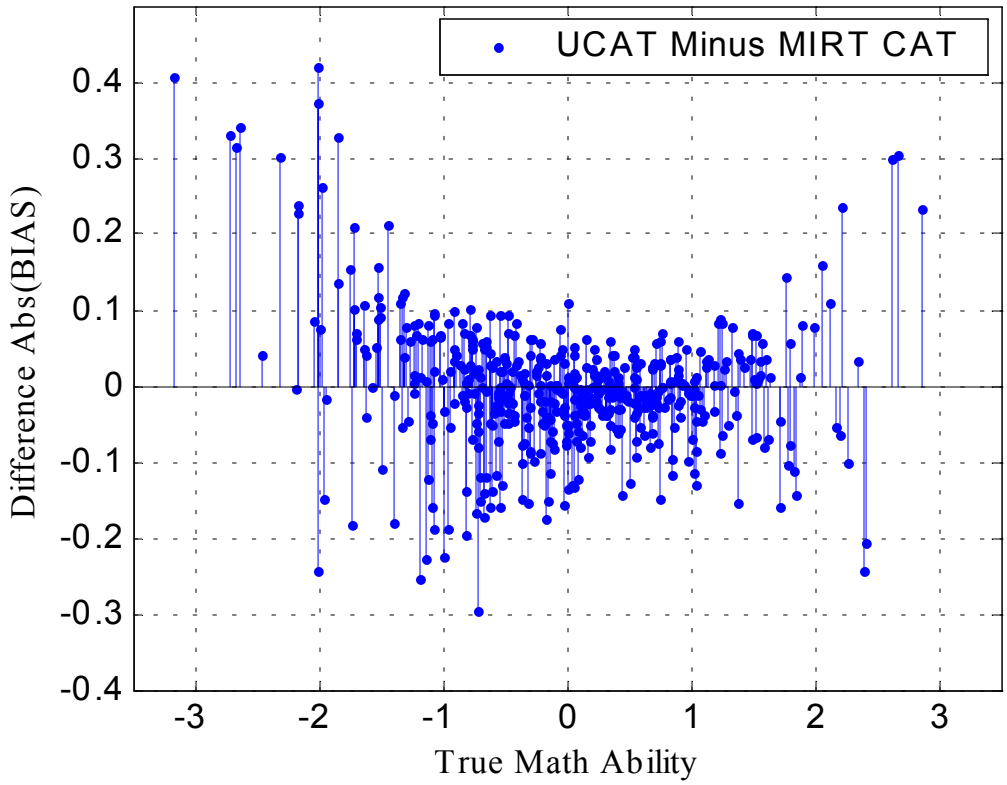

Figure 2b Difference Abs(BIAS) as a Function of True Math Ability

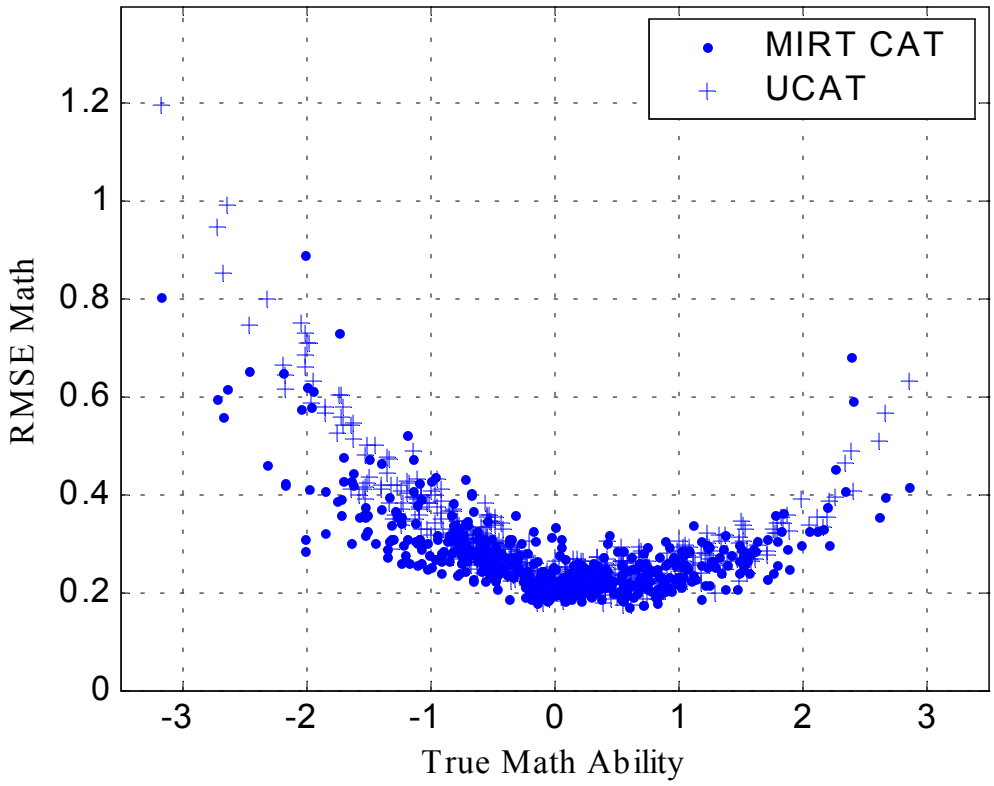

Figure 2c RMSE as a Function of True Math Ability for MIRT CAT and UCAT

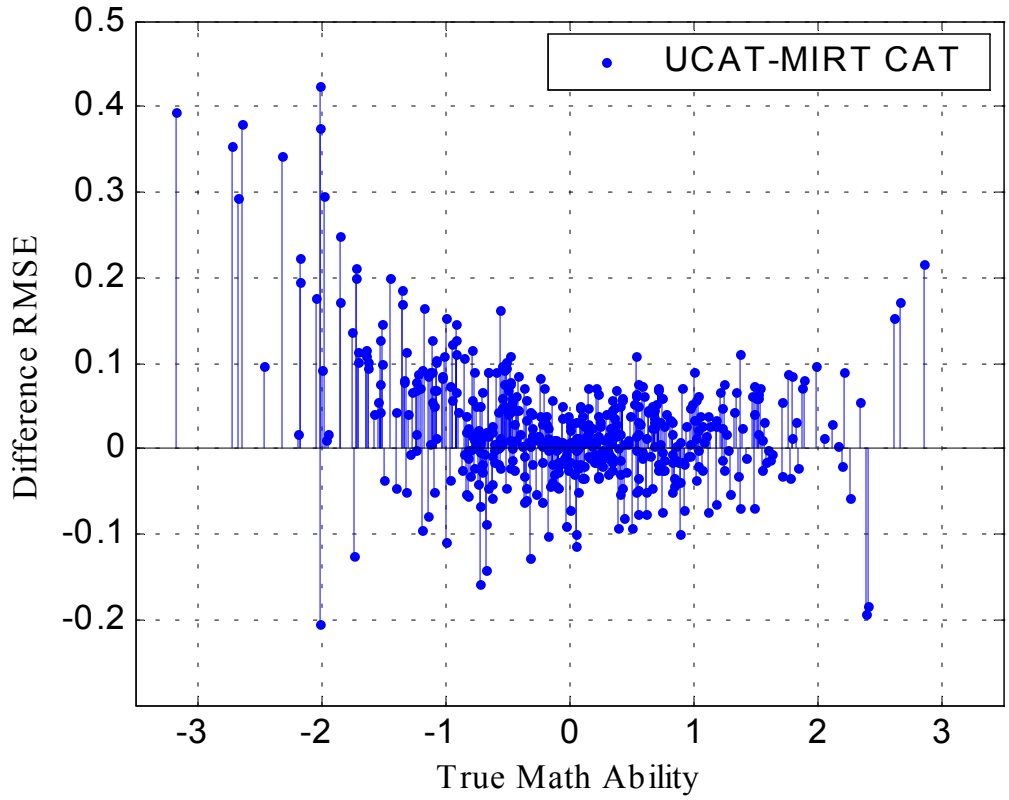

Figure 2d Difference RMSE as a Function of True Math Ability

In short, figures 1 and 2 combined with the above summary descriptive statistics suggest that MIRT CAT tends to increase the accuracy of ability estimates, especially for low or high Reading and Math ability estimates.

2. Comparison Between Reading and Math Ability Estimates

For the condition of the shadow-test MIRT CAT by maximizing the DPI, the standard deviation of BIAS across 500 simulees was .085 and .143 for Reading and Math, respectively. The average RMSE was .212 and .276 for Reading and Math, respectively. Similar summary statistics were found for the second research condition that employed the UCAT to the same multidimensional item pool.

As seen as Table 4, the range of RMSE for Reading ability estimates under MIRT CAT was between .121 and .515; in contrast, Math ability estimates exhibited a larger range of RMSE, from .170 to .891. These results suggest that Reading abilities were more accurately estimated than Math. The primary reason why these results occurred may be associated with the magnitude of the discrimination parameters. As Li and Lissitz (2000) pointed out, the accuracy of multidimensional ability estimates in each dimension relies on the corresponding magnitude of the discrimination parameters. The values of discrimination parameters in Math in the simulated item pool are relatively smaller than in Reading so that MIRT CAT faced more challenge in recovering the Math abilities than the Reading ones.

3. Intercorrelations Among True Reading, True Math and their Ability Estimates

The intercorrelations among the true Reading, the true Math and their ability estimates were calculated and presented in Table 5. The Pearson correlation between the true- and estimated abilities were .998 and .994 (Table 5) for Reading and Math, respectively for the MIRT CAT research condition. Similar correlation coefficients were found in the UCAT condition. The correlation between the true Reading and Math abilities across 500 simulees was .725; in contrast, its corresponding correlation computed from the recovered abilities was .798 in the MIRT CAT condition. This slightly inflation, also found in Segall study (1996), may be due to the bias in ability estimates introduced by the prior covariance of ability estimates. In contrast, this did not occur in the UCAT condition probably because the prior covariance of ability estimates was set at zero in the UCAT condition, in which the correlation computed from the recovered abilities was .718.

Table 5. Correlation Matrix for the True and Estimated Reading and Math Abilities

Contents	1	2	3	4	5	6
1. Reading (True)	1.000					
2. Math (True)	0.725	1.000				
3. MIRT CAT Reading	0.998	0.748	1.000			
4. MIRT CAT Math	0.777	0.994	0.798	1.000		
5. UCAT Reading	0.999	0.721	0.998	0.774	1.000	
6. UCAT Math	0.721	0.995	0.744	0.994	0.718	1.000

B. Item Exposure Rate

The results of item exposure rates are presented in Table 6. The plot of each item's exposure rate for MIRT CAT and UCAT methods are presented in Figure 3.

The rates of unused items are high in both research conditions, especially in the UCAT algorithm that had 62.22% of the items never being selected for any of the 5000 examinees and .13% items selected for all examinees. In contrast, the MIRT CAT algorithm used 5.62 % more of the pool, still leaving 56.6% of the item pool unused, and no item was selected for all examinees. The MIRT CAT algorithm's item exposure rates were slightly more homogeneous, as indicated by a smaller SD value (see Table 6). The results of item-exposure rates yielded in the current study may not be fully satisfactory for all practitioners. An additional item-exposure control (for details, see Revuelta and Ponsoda, 1998) might be needed to be incorporated into MIRT CAT if reducing the rate of unused items and increasing the homogeneity of used test items are desirable. This is another subject to be explored in the context of MIRT CAT.

In short, the item-exposure results combined with the results regarding the accuracy of MIRT CAT ability estimates suggests that MIRT CAT is the CAT algorithm of choice when compared with UCAT.

Table 6:
Item Exposure Rate Distribution for the MIRT CAT and UCAT methods

Item Exposure Rate (x100)	MIRT CAT	UCAT
0: Items Never Administered	56.60	62.22
1-5	17.13	13.46
5-10	5.23	4.05
10-15	4.05	3.53
15-20	2.61	2.48
20-25	3.66	3.66
25-30	2.75	2.61
30-35	1.83	1.70
35-40	1.83	1.44
40-45	1.44	1.44
45-50	0.92	1.05
50-60	1.57	1.18
60-70	0.13	0.78
70-80	0.00	0.00
80-90	0.00	0.00
90-99	0.26	.26
100: Items Always Administered	0.00	.13
Mean(x100)	6.67	6.67
SD(x100)	13.45	14.20
Minimum(x100)	.00	.00
Maximum(x100)	91.98	100.00

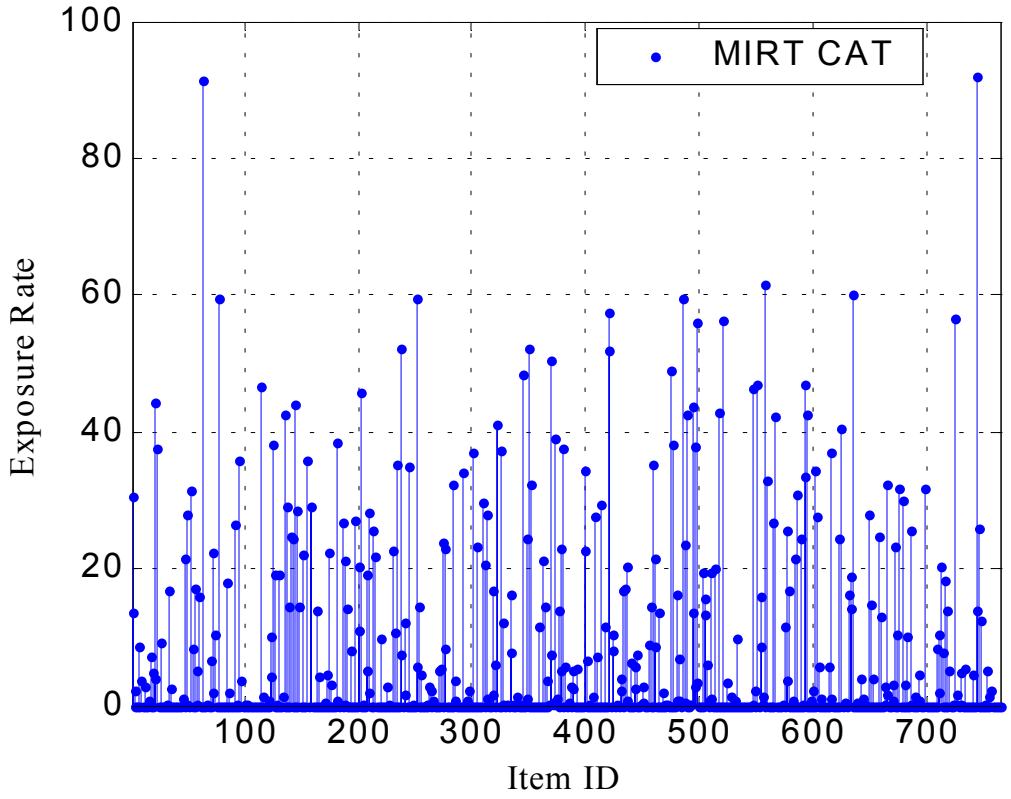

Figure 3a. Plot of Each Item's Exposure Rate for MIRT CAT Method

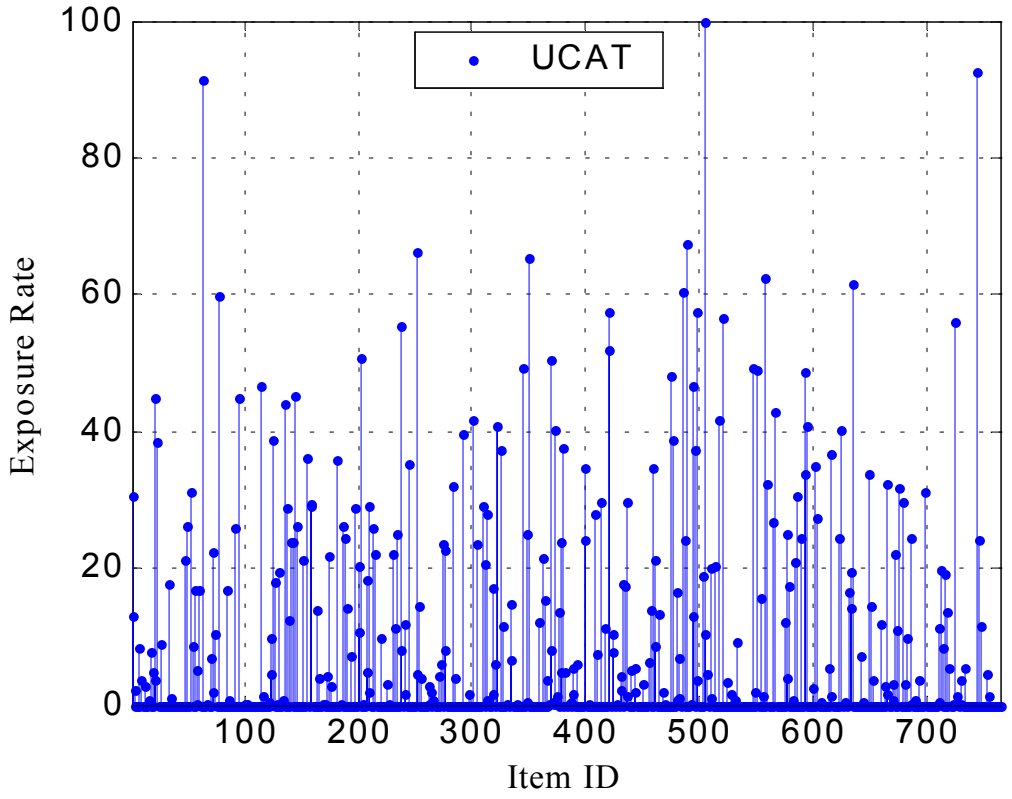

Figure 3b. Plot of Each Item's Exposure Rate for UCAT Method

V. Conclusions and Applications

The current goal of the MIRT CAT study is to implement it in a real testing program. To achieve this end, it may be better to begin with the application of MIRT CAT on a test battery with multiple content-area scores (e.g., Luecht, 1996 and Segall, 1996). This testing scenario is preferred at the current stage of MIRT development for two reasons: (1) Identifying the interpretable latent dimensions for this type of test data is relatively feasible, and (2) The report of multiple content-area scores as resulting by MIRT CAT will be more readily accepted by the public and very useful for instructional or other uses (e.g., Luecht, 1996 and Segall, 1996). On the other hand, if the reported test outcome is strictly limited to a single score, there might be little apparent advantage in considering a MIRT model that attempts to capture the salient multidimensionality of the test-examinee interaction (Luecht, 1996).

To create the testing scenario as described above, this study used two real tests of CTBS Reading and Math as the simulated tests. As indicated, a confirmatory item factor-analysis model has been implemented by the NOHARM to produce more interpretable MIRT item parameters. As a result, the simulated item parameters used in this study have two special features: (a) Reading or Math items contributed only its intended measures, and (b) since the simulated item discrimination parameters were intentionally defined for the Reading and Math dimensions, the MIRT CAT ability estimates should correspond to both measures, respectively, rather than to the abstract $\underline{\theta}$.

The study is expected to provide evidence to demonstrate that the algorithm of MIRT CAT, as presented in this study, can feasibly to be implemented in the testing scenario mentioned above. The results presented in the current study showed that MIRT CAT is quite capable of producing quite accurate estimates on each of the content-area scores using relatively short test length. The covariance between Reading and Math facilitated a better MIRT CAT item selection strategy than if both tests are separately administered using UCAT. The current large-scale testing programs (e.g., SAT, ACT, GRE, GMAT, LSAT, etc.) usually use more test length than that used in this study. If they employ the MIRT CAT as implemented in this study, it is anticipated that the accuracy of ability estimates will increase, especially for the low or high ability estimates.

The simulees took 25 Reading items first and 26 Math items next under the MIRT CAT implemented in the current study. This setting might improve the efficiency of MIRT CAT, especially for the Math measure, for two reasons: (a) the first-stage testing on the Reading measure provided some information of the Math ability estimate that is valuable to be used as a starting value when the second-stage Math measure begins, and (b) during the second-stage Math measure, the Reading ability estimates are themselves updated and a more accurate estimate can be obtained. Based on this analogy, the MIRT CAT will do much better if it were implemented in a testing setting that requires simultaneously estimate multiple abilities (more than two) for each of examinees.

Seeking an item with the maximum DPI can rapidly improve the ability estimates. As noted, an item's DPI value is a function of an ability estimate. However, at the beginning of MIRT CAT, ability estimates are not as accurate as those estimated at the latter stage and consequently the item's DPI value at the earlier stage has small impact on the final-stage ability estimates. It appears that the DPI criterion in item selection is not very important until better ability estimates are obtained.

The results of this study show that MIRT CAT with a shadow-test algorithm did not result in very satisfactory result in terms of item-exposure rates. As indicated, items in the on-line assembling shadow test have much larger DPI values than the rest of the items in the pool and the earlier-stage DPI values are not as meaningful as previously defined (Segall, 1996, 2000). Practitioners can make use of these facts in designing a better approach in item selection at the beginning stage of MIRT CAT under the shadow test context. For example, if we randomly select test items from the shadow test at the first and the second quarter test, the goal of item exposure control will be better achieved and perhaps this approach to item selection will not sacrifice the test precision. This proposed method (or other methods, illustrated next) can be examined in assessing its impact on ability estimates.

There are two key components for the shadow-test MIRT CAT. One is the information (or DPI) function to be maximized; the other is the constraints imposed by users. At the beginning of the shadow-test CAT (e.g., the first two quarter tests), we are maximizing something (e.g., information or DPI) that may not be as meaningful as we anticipate. We may be emphasizing information (or DPI in the MIRT CAT case), technically by maximizing (or minimizing) other functions, while the on-line shadow test is assembled at the beginning of the CAT (or MIRT CAT). A modification that allocates more emphasis to user constraints designed to minimize exposure may be possible. Such a modification might dramatically expand the use of all characteristics of items in the pool and guarantee that all constraints are fully met. What other target (or objective) functions are legitimate (or appropriate) to be maximized (or minimizing) at the beginning of CAT test? Part of this issue has been addressed in Li and

Schafer's (2003) study in which UCAT was used. More future MIRT CAT studies might help to address this issue. When more MIRT CAT studies are conducted, the special features of MIRT and MIRT CAT could then be utilized in real testing situations.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 18, 255-278.

Baker, F. B. (1992). Item Response Theory: Parameter Estimation Techniques. New York: Marcel Dekker, Inc.

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

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Boxom, B. M., & Vale, C. D. (1987, June). Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta. Paper presented at the meeting of the Psychometric Society, Montreal.

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

CTBS/McGraw-Hill (1997). Teacher's guide to TerraNova. Monterey, CA. McGraw-Hill Companies, Inc.

Educational Testing Service. (1993). The GRE computer adaptive testing program (CAT): Integrating convenience, assessment, and technology. Princeton, NJ: Educational Testing Service

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

Gibbons, D. D., & Hedeker, D. R. (1992). Full information item bi-factor analysis. Psychometrika, 57, 423-436.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. Applied Psychological Measurement, 24, 115-138.

Li, Y. H., & Schafer, W. D. (2003). Increasing the homogeneity of CAT's Item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

LINDO Systems, Inc. (2001). LINDO API: The premier optimization engine. [Computer program]. Chicago Illinois: INDO Systems, Inc.

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

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. Applied Psychological Measurement, 20, 389-404.

McLeond, L.D., Swygert, K. A, & Thissen, D. (2001). Factor analysis for items scored in two categories D. Thissen and H. Wainer (eds.), Test Scoring, 189-216.Mahwah NJ: Lawrence Erlbaum Associates, Inc.

Mislevy, R. J., & Bock, R. D. (1990). BILOG-3 (2nd ed.): Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

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.

Rao, P. S. (2000). Sampling methodologies with applications. New York: Chapman & Hall/CRC.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.
Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 271-286). New York: Springer.
Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. Journal of Educational Measurement, 35, 311-327.
Segall, D. O. (1996). Multidimensional adaptive testing. Psychometrika, 61, 331-354.
Segall, D. O. (1998). IFACT computer program Version 1.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation [computer program]. Seaside, CA: Defense Manpower Data Center.
Segall, D. O. (2000). Principles of Multidimensional Adaptive Testing. W. J. van der Linden and C. A. W. Glas (eds.), Computerized Adaptive Testing: Theory and practice, 53-57. Dordrecht, The Netherlands: Kluwer Academic Publishers.
Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. Psychometrika, 66, 79-97.
Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. Proceeding of the 27th Annual Meeting of the Military Testing Association (pp. 937-977). San Diego, CA: Navy Personnel Research and Development Center.
Tam, S. S. (1992). A comparison of method for adaptive estimation of a multidimensional trait. Unpublished doctoral dissertation, Columbia University, New York City, NY.
The MathWorks, Inc. (2001). MATLAB (Version 6.1): The language of technical computing [Computer program]. Natick MA: The MathWorks, Inc.
Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.
Theunissen, T. J. J. M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, 381-389.
van der Linden (1998). Optimal assembly of psychological and educational tests. Applied Psychological Measurement, 22, 195-211.
van der Linden, W. J. (1999). Multidimensional Adaptive testing with a minimum error-variance criterion. Journal of Educational and Behavioral Statistics, 24, 398-412.
van der Linden, W. J. (2000). Constrained Adaptive Testing with Shadow Tests. W. J. van der Linden and C. A. W. Glas (eds.), Computerized Adaptive Testing: Theory and practice, 27-52. Dordrecht, The Netherlands: Kluwer Academic Publishers.
van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.
van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. Applied Psychological Measurement, 22, 259-270.
Veldkamp, B. & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content, Psychometrika, 67, 575-588.
Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content, Psychometrika, 67, 575-588.
Wood, R., Wilson, D., Gibbons, R, Schilling, S., Muraki, E., & Bock, D. (2003). TESTFACT 4: Test scoring, item statistics, and item factor analysis. Mooresville, IN: Scientific Software.

Appendix A. Item Classification Table for CTBS, Grade 2 Test

Reading Domain	**Objective and Skills**	**Number of Items**
1 Basic Understanding	1 vocabulary	3
	2 stated information	5
2 Analyze Text	3 conclusions	4
	4 story elements/character	3
3 Evaluate and Extend Meaning	5 predict/hypothesize	1
	6 extend/apply meaning	3
4 Identify Reading Strategies	7 make connections	2
	8 vocabulary strategies	4
Total for Reading Test		25
Math Domain		
1. Number and Number Relations	1 counting	2
	2 compare, order	1
	3 ordinal numbers	1
	4 fractional part	1
	5 place value	1
2. Computation and Numerical Estimation	6 computation	3
	7 computation in context	1
	8 computation with money	1
3 Operation Concepts	9 model problem situation	2
4 Measurement	10 time	1
	11 use ruler	1
5 Geometry and Spatial Sense	12 solid figure	1
	13 congruence, similarity	1
	14 symmetry	1
6 Data Analysis, Statistics and Probability	15 read bar graph	1
	16 probability	1
	17 use data to solve problems	1
	18 compare data	1
7 Patterns, Functions, Algebra	19 number pattern	1
8 Problem Solving	20 solve nonroutine problem	1
	21 proportional reasoning	2
Total for Math Test		26