Asia Pacific Education Review
2007, Vol. 8, No. 1, 41-55.

Evaluation of Linking Methods for Multidimensional IRT Calibrations

Kyung-Seok Min
Korea Institute of Curriculum & Evaluation
Korea

Most researchers agree that psychological/educational tests are sensitive to multiple traits, implying the need for a multidimensional item response theory (MIRT). One limitation of applying a MIRT in practice is the difficulty in establishing equivalent scales of multiple traits. In this study, a new MIRT linking method was proposed and evaluated by comparison with two existing methods. The results showed that the new method was more acceptable in transforming item parameters and maintaining dimensional structures. Limitations and cautions in using multidimensional linking techniques were also discussed.

Key words: linking, multidimensional item response theory, orthogonal Procrustes rotation, scale indeterminacy.

Introduction

Compared with traditional classical test theory, the most important characteristic of item response theory (IRT) is the invariance of item parameters. That is, item difficulty and discrimination remain unchanged across different examinee groups (Lord, 1980). However, IRT models require more demanding statistical assumptions such as unidimensionality and local independence.

An IRT framework has been developed under the assumption of unidimensionality; the item-person interaction is modeled with a single latent trait. However, the mechanisms and cognitive processes that an examinee uses to respond to test items do not always appear to be so simple, and many psychological and educational researchers agree that multidimensional abilities/traits come into play in test performance (Ackerman, 1996; Reckase, 1995). For example, in order to solve a math item with verbal descriptions, examinees must not only know about calculations but also understand related verbal descriptions. Therefore, this math item is supposed to measure both the math and verbal ability of examinees.

On the other hand, most testing programs administer different test forms on different days because of test security and flexibility, but test scores of different administrations should be interchangeable to provide consistent score information. Even though test developers attempt to make test forms similar, nevertheless, the forms typically differ somewhat. Therefore, statistical procedures such as linking or equating are needed to adjust for different levels of difficulty across test forms.

Most IRT linking methods have been based on unidimensional item response theory (UIRT), and UIRT linking makes adjustments for different scales (i.e., origin and unit of scale) (Lord, 1980). When the goal is to establish comparable scales on tests that are affected by more than one dimension, however, the directions of dimensions also need to be adjusted to obtain equivalent scales. That is, multidimensional item response theory (MIRT) models are

Kyung-Seok Min, Office for the College Scholastic Ability Test, Korea Institute of Curriculum & Evaluation.
This work was supported partly by the College Board (#2003-123). I am grateful to Dr. Mark D. Reckase at Michigan State University and Dr. Vytas Laitusis at the College Board for offering helpful suggestions at many important points. Any errors are my own.
Correspondence concerning this paper should be addressed to Korea Institute of Curriculum & Evaluation, 25-1, Samchung-dong, Jongno-gu, Seoul, 110-230, Korea. e-mail: minks@kice.re.kr.

directionally indeterminate as well as scale indeterminate. Therefore, MIRT linking requires a composite transformation of rotation and scaling to derive comparable scores (Li & Lissitz, 2000).

Even though most psychological and educational tests are sensitive to multiple traits/skills, implying the need for MIRT, the application of MIRT is limited in practice by difficulties in establishing equivalent scores on multiple ability dimensions. While several MIRT linking methods have already been developed to solve the problem of comparability (e.g., Hirsch, 1989; Li & Lissitz, 2000; Oshima, Davey, & Lee, 2000; Thompson, Nering, & Davey, 1997), each has unique properties in terms of statistical characteristics and optimization criteria. Moreover, it is not yet known whether different MIRT linking methods lead to the same/similar metric transformations even though there has been one comparison study (Min & Kim, 2003) so far.

Purpose of the Study

The purpose of this study is to propose a new linking method that can provide more desirable multidimensional metric transformations, especially in the dilation/contraction of a scale, and evaluate the new method by comparing it with two existing linking procedures for MIRT scales (i.e., Li & Lissitz, 2000; Oshima et al., 2000) in terms of the accuracy and stability of metric transformations under a number of testing conditions.

The MIRT linking method developed by Li and Lissitz (2000) includes only a single dilation constant for multiple dimensions based on traditional factor analysis techniques (i.e., orthogonal Procrustes solutions, Schönemann & Carroll, 1970). However, more desirable transformations can possibly be expected when linking allows a unique dilation/contraction for each dimension. A new MIRT linking method that incorporates a diagonal dilation matrix into orthogonal Procrustes solutions was proposed and compared with the previous two linking methods.

MIRT Models and Linking Methods

MIRT Models

Two types of MIRT models have been developed, compensatory (Reckase, 1995) and noncompensatory models (Sympson, 1978). Since most research on MIRT has been done using compensatory models (partly because of estimation difficulties for noncompensatory models), and the fit of the two types of MIRT models appears indistinguishable from a practical point of view, the compensatory model is considered in this study.

The compensatory multidimensional extension of the three-parameter logistic model with m dimensions is (Reckase, 1995)

$$P(u_{ij}=1 \mid \mathbf{a}_i, c_i, d_i, \boldsymbol{\theta}_j) = c_i + (1-c_i)\frac{\exp(\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)}{1+\exp(\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)}, \quad (1)$$

where $P(u_{ij}=1 \mid \mathbf{a}_i, c_i, d_i, \boldsymbol{\theta}_j)$ is the probability of a correct response for examinee j on test item i in an m-dimensional space, u_{ij} is the item response for person j on item i (1 correct; 0 wrong), $\mathbf{a}_i$ is a vector of discrimination parameters of item i, c_i is the lower asymptote (probability of correct answer when an examinee's ability is very low), d_i is a parameter related to item difficulty of item i, and $\boldsymbol{\theta}_j$ is a vector of the jth examinee's abilities.

The MIRT difficulty and discrimination factors are not directly equivalent to those of UIRT because of different parameterizations. Two statistics are used to capture multidimensional item characteristics corresponding to unidimensional item discrimination and difficulty. The discrimination power of a multidimensional item is (Reckase & McKinley, 1991)

$$MDISC_i = \left(\sum_{k=1}^{m} a_{ik}^2 \right)^{1/2}, \quad (2)$$

where $MDISC_i$ denotes the ith item's multidimensional discrimination as a function of the slope at the steepest point, and a_{ik} is the ith item's discrimination on the kth dimension.

Multidimensional item difficulty equivalent to unidimensional difficulty is

$$MDIFF_i = \frac{-d_i}{MDISC_i}, \quad (3)$$

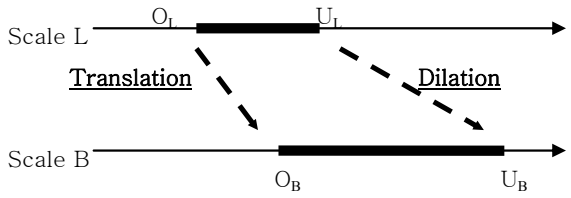
where $MDIFF_i$ is the distance between the origin and the point of the steepest slope on the ability space.

The direction of the greatest discrimination in the dimensional space is given by

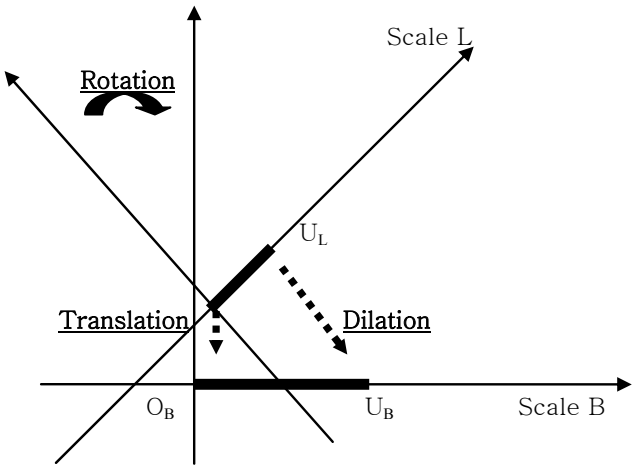
$$\alpha_{ik} = \arccos\frac{a_{ik}}{MDISC_i} \quad (\text{or } \cos\alpha_{ik} = \frac{a_{ik}}{MDISC_i}), \quad (4)$$

where α_{ik} is an angle from the kth dimension.

As is shown in Equation 1, the probability of the correct



(a) UIRT Linking



(b) MIRT Linking (two-dimension case)

Figure 1. UIRT and MIRT Linking Components. *O* is the location of origin, *U* is the length of unit, the subscript *L* is the metric to transform and *B* is the target/base metric. From "An evaluation of the accuracy of multidimensional IRT equating," by Y. H. Li and R. W. Lissitz, 2000, *Applied Psychological Measurement*, *24*, p. 120. Copyright 2000 by SAGE Publications. Adapted with permission.

answer is a linear function of item ($\mathbf{a}$ and d) and ability

($\boldsymbol{\theta}$)parameters in the exponent. Therefore, any linear transformation of an ability scale results in the same value of the exponent for a given response pattern if the item and ability parameters are transformed in a consistent way. While scale indeterminacy is a concern when finding a proper transformation in UIRT linking (translation of the origin and dilation of the unit length in the upper part of Figure 1), the rotation to determine the comparable reference system as well as the scale alterations has to be considered in MIRT due to the issue of multidimensionality (translation, dilation and rotation of the reference system in the lower part of Figure 1).

MIRT Linking Methods

So far, several MIRT linking methods have been proposed (e.g., Hirsch, 1989; Li & Lissitz, 2000; Oshima et al., 2000; Thompson et al., 1997). These methods use a two-dimensional compensatory model and consist of some or all of three linking components; a rotation matrix dealing with directional indeterminacy, and a translation vector and a dilation constant which removes scale indeterminacy of the origin and unit. Hirsch's study (1989) is valuable since it is the first attempt to deal with multidimensionality in IRT linking, and his procedures are expanded in Li and Lissitz' method later. Additionally, the method of Thompson et al. (1997) has great potential, but it is still experimental. Therefore, the focus of this study is on the two most recent MIRT linking methods (i.e., Li & Lissitz, 2000; Oshima et al., 2000) to compare with the new method.

Oshima, Davey, and Lee's method: TCF method. Oshima and her colleagues' linking method (2000) is based on the anchor item design; in which a set of common items are included in multiple test forms to define common scales. Transformations of the parameters of the compensatory model with the exponent of $\mathbf{a}_i' \boldsymbol{\theta}_j + d_i$, are conducted through the following equations.

$$\mathbf{a}_i^* = (\mathbf{A}^{-1})'\mathbf{a}_i, \tag{5}$$

$$d_i^* = d_i - \mathbf{a}_i' \mathbf{A}^{-1}\boldsymbol{\beta}, \tag{6}$$

$$\boldsymbol{\theta}_j^* = \mathbf{A}\boldsymbol{\theta}_j + \boldsymbol{\beta}, \tag{7}$$

where $\mathbf{A}$ ($m \times m$, m is the number of dimensions) is a rotation matrix, $\boldsymbol{\beta}$ ($m \times 1$) is a translation vector, and the asterisk (*) indicates transformed parameters. Here, the rotation matrix $\mathbf{A}$ has two functions, (a) rotate to a proper dimensional orientation, and (b) adjust the variances of ability dimensions. The translation vector $\boldsymbol{\beta}$ is used to shift to a compatible origin by altering the origin of a scale. Once $\mathbf{A}$ and $\boldsymbol{\beta}$ are identified, then a_i^* , d_i^* , and $\boldsymbol{\theta}_j^*$ are in order.

The equality of the transformed exponent and the original one can be illustrated by

$$\mathbf{a}_i'^{*}\boldsymbol{\theta}_j^* + d_i^* = (\mathbf{a}_i' \mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\theta}_j + \boldsymbol{\beta}) + (d_i - \mathbf{a}_i' \mathbf{A}^{-1}\boldsymbol{\beta}) =$$

$$\mathbf{a}_i' \boldsymbol{\theta}_j + d_i. \tag{8}$$

Oshima et al. compared four estimation procedures and concluded that the test characteristic function method (TCF) was best at estimating the rotation matrix and was also relatively effective in estimating the translation vector. The minimization function for the TCF method is

$$T(\boldsymbol{\theta}) = \sum_{i=1}^{n} P_i(\boldsymbol{\theta}) \ ,$$
$$\sum_{\boldsymbol{\theta}} w_{\boldsymbol{\theta}} [T_B(\boldsymbol{\theta}) - T_L^*(\boldsymbol{\theta})]^2 \ , \quad (9)$$

where T_B and T_L^* indicate expected number-correct scores (i.e., true scores) for the common items on the base test and the linked test, respectively, n is the number of common items, and $w_{\boldsymbol{\theta}}$ is a weighting value (e.g., inverse of measurement errors or other reasonable numbers) which allows some regions on the ability space of $\boldsymbol{\theta}$ to be more important than others. If weights are equal for all regions, the result is an unweighted estimation. Equation 9 indicates that $\mathbf{A}$ and $\boldsymbol{\beta}$ can be identified by minimizing the gaps between the TCFs of the base test and the linked test. Therefore, the TCF method is a multidimensional extension of Stocking and Lord's method (1983) that develops a common metric by minimizing differences of item characteristic curves.

Li and Lissitz' method: Trace method. Li and Lissitz (2000) developed four different linking procedures based on the anchor item design and claimed that the best procedure was a composite transformation with three components; a rotation matrix from the orthogonal Procrustes solutions, a translation vector obtained by a least-square criterion, and a central dilation constant obtained by the trace method. In order to emphasize the difference of the dilation component from the new method, which will be described later, Li and Lissitz' method will be referred to as the Trace method.

The Trace method uses the following equations to transform model parameters in the exponent, $\mathbf{a}_i' \ \boldsymbol{\theta}_i + d_i$,

$$\mathbf{a}_i^* = k\mathbf{a}_i' \ \mathbf{T} \ , \quad (10)$$
$$d_i^* = d_i - \mathbf{a}_i' \ \mathbf{Tm} \ , \quad (11)$$
$$\boldsymbol{\theta}_j^* = (1/k)(\mathbf{T}^{-1}\boldsymbol{\theta}_j + \mathbf{m}) \ , \quad (12)$$

where $\mathbf{T}$ (m × m) is an orthogonal rotation matrix, $\mathbf{m}$ (m × 1) is a translation vector for location, and k (1 × 1) is a central dilation constant for unit change. Then the equality of exponent terms after and before transformation is established by

$$\mathbf{a}_i^* \boldsymbol{\theta}_j^* + d_i^* = (k\mathbf{a}_i' \ \mathbf{T})(1/k)(\mathbf{T}^{-1}\boldsymbol{\theta}_j + \mathbf{m}) + (d_i - \mathbf{a}_i' \ \mathbf{Tm})$$
$$= \mathbf{a}_i' \ \boldsymbol{\theta}_j + d_i \ . \quad (13)$$

The linking components of the Trace method are obtained by minimizing the following functions.

$$\mathbf{E}_1 = k\mathbf{A}_L\mathbf{T} - \mathbf{A}_B \ , \quad (14)$$
$$tr(\mathbf{E}_1'\mathbf{E}_1) = tr\{(k\mathbf{A}_L\mathbf{T} - \mathbf{A}_B)'(k\mathbf{A}_L\mathbf{T} - \mathbf{A}_B)\} \ , \quad (15)$$
$$Q = \sum_{i=1}^{n} (d_{iB} - d_{iL}^*)^2 \ . \quad (16)$$

In Equations 14 to 16, *tr* is the matrix operator of the sum of diagonal elements (trace), $\mathbf{A}$ (n × m) and d_i are discrimination matrix and difficulty-related parameter.

Extension of the Trace method with a diagonal dilation matrix: DDM method. Li and Lissitz clearly indicated that they assumed a constant change of the unit lengths across multiple dimensions such that one dilation constant was enough to cover overall unit length adjustments. They provided two reasons for a dilation constant: mathematical tractability and relatively reasonable accuracy. In the TCF method, this issue was not clearly stated, however, the simulation examples (Oshima et al., 2000, Table 2 on p. 364) showed that their main concern was about a constant unit change across dimensions. One reasonable argument for constant overall dilation of multiple dimensions may be that the dimensions measured by a test are strongly related, such that the change in one dimension goes along with other dimension(s) at the same dilation/contraction rate. However, this may not be typical for various constructs measured by educational/psychological tests. In addition, from a methodological perspective, the dilation constant can be treated as a special case of multiple dilation constants.

In order to model different unit changes along with an orthogonal rotation, the dilation constant adopted in the Trace method is replaced with a *d*iagonal *d*ilation *m*atrix, referred to as the DDM method, to emphasize the difference of the dilation component from the Trace method. Transformation equations of the DDM method are

$$\mathbf{a}_i^* = \mathbf{a}_i' \ \mathbf{TK} \ , \quad (17)$$
$$d_i^* = d_i - \mathbf{a}_i' \ \mathbf{Tm} \ , \quad (18)$$
$$\boldsymbol{\theta}_j^* = \mathbf{K}^{-1}(\mathbf{T}^{-1}\boldsymbol{\theta}_j + \mathbf{m}) \ , \quad (19)$$

where $\mathbf{K}$ is a diagonal dilation matrix and other terms are defined as before. For a two-dimension case, $\mathbf{K}$ is defined as $\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$ here, k_1 indicates the dilation component for the first dimension, and k_2 is for the second

dimension. Off-diagonal elements of **K** are set to zero because the relationship/direction between two dimensions is not defined by **K** but only by the orthogonal rotation matrix, **T**. In this case, the equality of exponent terms is established by

$$\mathbf{a}_i^{*}\boldsymbol{\theta}_i^{*} + d_i^{*} = (\mathbf{a}_i' \ \mathbf{TK})(\mathbf{K}^{-1})(\mathbf{T}^{-1}\boldsymbol{\theta}_j + \mathbf{m}) + (d_i - \mathbf{a}_i' \ \mathbf{Tm})$$
$$= \mathbf{a}_i' \ \boldsymbol{\theta}_j + d_i . \qquad (20)$$

Two points should be mentioned. First, Equations 17 to 19 are the same as Equations 10 to 12 except for including a diagonal dilation matrix rather than a dilation constant. Second, when all diagonal components of **K** are equal, Equation 20 becomes the same as Equation 13. The proposed linking method in Equations 17 to 19 differs from the TCF method by splitting the rotation matrix and the dilation matrix, and using an orthogonal rotation. It also differs from the Trace method by allowing a unique unit change for each dimension rather than a constant change for all dimensions.

Two minimization criteria for the DDM method for the rotation matrix and the diagonal dilation matrix are provided in Equations 21 and 22, respectively (once **T** is identified, then **K** is in order), and the criterion of the translation vector is the same as the Trace method, Equation 16.

$$tr(\mathbf{E}_2'\mathbf{E}_2), \text{ where } \quad \mathbf{E}_2 = \mathbf{A}_L\mathbf{T} - \mathbf{A}_B \qquad (21)$$

$$\mathbf{E}_3'\mathbf{E}_3, \text{ where } \mathbf{E}_3 = \mathbf{A}_L\mathbf{TK} - \mathbf{A}_B \qquad (22)$$

Method and Analysis

Simulation Data Analysis

It is recommended to use simulation data to evaluate linking methods in order to separate the effect of model misfit and linking errors (Harris & Crouse, 1993). Since we know the true parameters in the simulation study, it is easier to compare true parameters with their estimates. Two test forms that share a set of common items were considered, the so-called common item design. Suppose one was the base test form and the other was the linked test form, and each form included common items and unique items. In such a case, the linked test scores need to be converted into the base test scores. The common item set consisted of 20 items for both tests, and they were used as a way of discovering a comparable test scale. Since the main purpose of this study is to find a common metric across different testing conditions rather than final equating results, noncommon items were not considered. In order to calibrate item and ability estimates, a compensatory, two-dimensional two-parameter logistic model was used as in Equation 1 with all $c_i = 0$.

Item parameters and test response patterns. Item parameters were drawn from probability distributions as to which ranges were determined by the specification of dimensional structures. Two types of dimensional structures were investigated; approximate simple structure (APSS) and mixed structure (MS) (Roussos, Stout, & Marden, 1998). For the present simulation, an APSS was constructed using two sets of items (ten items for each). One set of items mainly loaded on the first dimension and the other set on the second dimension. In MS, there were four sets of items (five items for each). Two sets of items loaded heavily on one of the dimensions, and the remaining two sets were sensitive to composites of the two dimensions. To construct dimensional structures, angles (α) between item vectors and the first dimension were randomly drawn from a uniform distribution with given ranges of the dimensional structures.

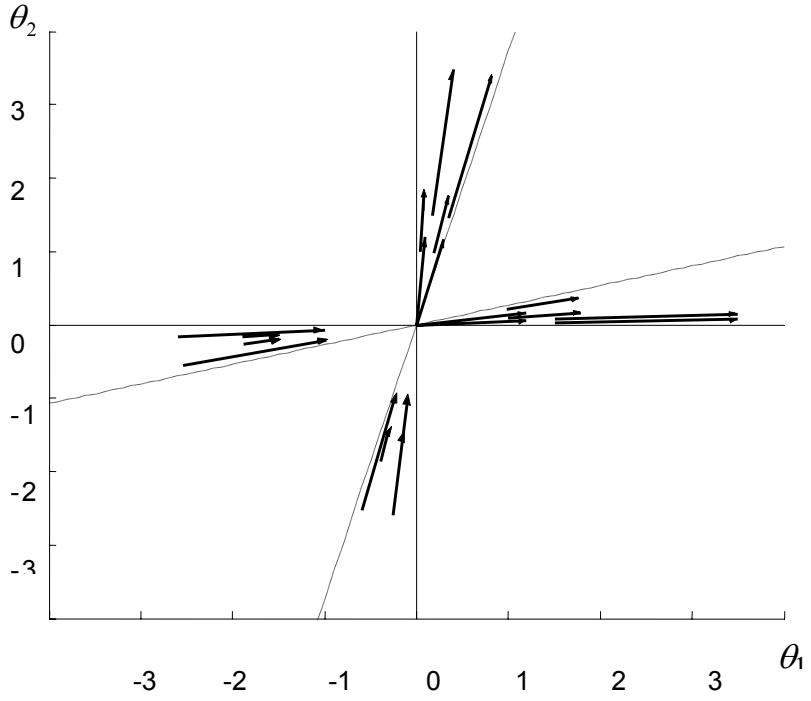
In order to define item parameters, fixed values of MDISCs and MDIFFs generated by Roussos et al. (1998) were used. The average value for MDISC is 1.2 (0.4, 0.8, 1.2, 1.6, & 2.0), and average value of MDIFF is zero (-1.5, 1.0, 0.0, -1.0, & 1.5). This pattern was repeated four times for 20 items. Discrimination and difficulty-related parameters were determined by Equations 2, 3, and 4 with given angles, MDISCs, and MDIFFs. The set of item parameters that were used for the present simulation is shown in Table 1.

For a visual presentation, directional vectors of twenty items are illustrated in Figure 3. APSS items in the upper part of Figure 3 were highly loaded on either dimension while MS items were widely spread between two dimensions. Therefore, an APSS implies a relatively independent dimensional structure and MS items tend to measure some combinations of two dimensions. The length of an item vector indicates the degree of discrimination (MDISC) and the distance between the origin and the starting point of the vector (arrow point of the vector on the third quadrant) is item difficulty (MDIFF). All vectors are extended through the origin, and they are located in the first and third quadrants because of positive discrimination parameters (a's) (Ackerman, 1996; Reckase & McKinley, 1991).

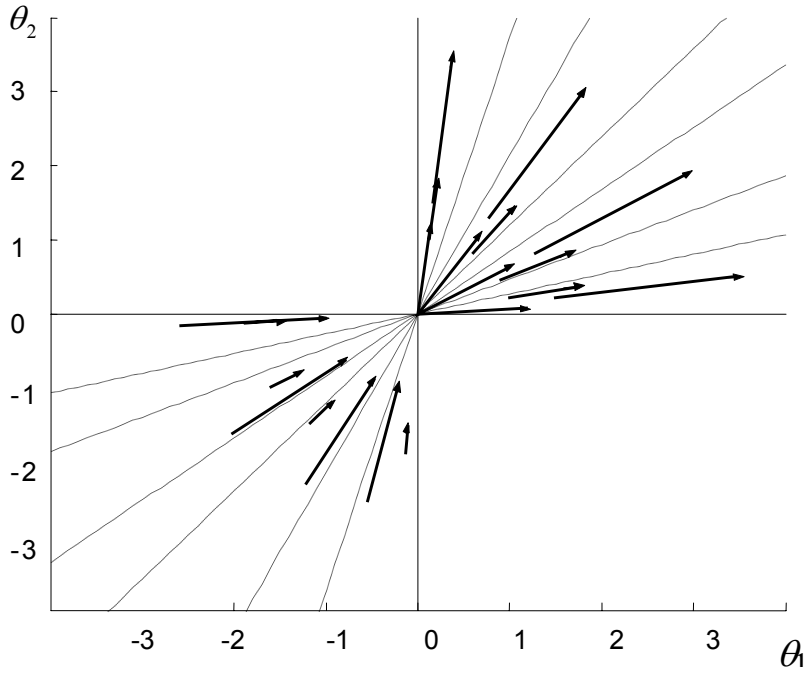
Finally, the probability of getting an item correct was computed by means of a two-dimensional IRT model, and this probability was compared with a random value drawn from a uniform distribution to generate dichotomous item response patterns (1 or 0).

Table 1. *Item Parameters of 20 Items*

Item	APSS		MS				
	a_1	a_2	a_1	a_2	d	MDISC	MDIFF
1	0.40	0.03	0.40	0.03	0.60	0.40	-1.50
2	0.80	0.07	0.78	0.17	-0.80	0.80	1.00
3	1.19	0.16	1.20	0.07	0.00	1.20	0.00
4	1.56	0.34	1.60	0.10	1.60	1.60	-1.00
5	2.00	0.04	1.98	0.29	-3.00	2.00	1.50
6	0.40	0.05	0.34	0.21	0.60	0.40	-1.50
7	0.78	0.17	0.71	0.36	-0.80	0.80	1.00
8	1.20	0.06	1.01	0.64	0.00	1.20	0.00
9	1.60	0.11	1.25	1.00	1.60	1.60	-1.00
10	2.00	0.09	1.68	1.08	-3.00	2.00	1.50
11	0.04	0.40	0.25	0.31	0.60	0.40	-1.50
12	0.15	0.79	0.47	0.65	-0.80	0.80	1.00
13	0.09	1.20	0.64	1.01	0.00	1.20	0.00
14	0.16	1.59	0.75	1.41	1.60	1.60	-1.00
15	0.47	1.94	1.03	1.71	-3.00	2.00	1.50
16	0.08	0.39	0.03	0.40	0.60	0.40	-1.50
17	0.04	0.80	0.10	0.79	-0.80	0.80	1.00
18	0.30	1.16	0.14	1.19	0.00	1.20	0.00
19	0.37	1.56	0.34	1.56	1.60	1.60	-1.00
20	0.23	1.99	0.21	1.99	-3.00	2.00	1.50
Mean	0.69	0.65	0.75	0.75	-0.32	1.20	0.00
SD	0.66	0.68	0.57	0.60	1.59	0.58	1.17



(a) Approximate Simple Structure (APSS)



(b) Mixed Structure (MS)

Figure 3. Item Vectors and Dimensional Structures

Table 2. *Ability Distributions of Five Examinee Groups*

Group 1	Group 2	Group 3	Group 4	Group 5
$\boldsymbol{\mu}=\begin{bmatrix}0\\0\end{bmatrix}$, $\boldsymbol{\Sigma}=\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	$\begin{bmatrix}0\\0\end{bmatrix}$, $\begin{bmatrix}1 & .5\\.5 & 1\end{bmatrix}$	$\begin{bmatrix}.5\\.5\end{bmatrix}$, $\begin{bmatrix}1 & .5\\.5 & 1\end{bmatrix}$	$\begin{bmatrix}.5\\.5\end{bmatrix}$, $\begin{bmatrix}.8 & .4\\.4 & .8\end{bmatrix}$	$\begin{bmatrix}.5\\.5\end{bmatrix}$, $\begin{bmatrix}1.2 & .5\\.5 & .8\end{bmatrix}$
ρ =0.00	0.50	0.50	0.50	0.51

Ability distributions. Five bivariate normal distributions with various means and variances/covariances were considered for illustrating the true abilities of examinees. Different ability distributions mean that the test forms were administered to somewhat different populations. Table 2 shows mean vectors ($\boldsymbol{\mu}$), variance/covariance matrices ($\boldsymbol{\Sigma}$) and correlation coefficients (ρ) of five examinee groups.

The distribution of group 1 is the default ability distribution assumed by the MIRT calibration program (e.g., NOHARM, Normal Ogive Harmonic Analysis Robust Method, Fraser, undated). Therefore, the smallest estimation and linking errors were expected for group 1. From group 2 to group 5 there were variations of mean vectors and variance/covariance matrices that require dimensional rotation and scaling to find common metrics.

Number of examinees. Usually 2000 or more examinees have been recommended for MIRT calibration. In order to evaluate the stability of linking, a relatively small number of examinees (1000) along with the recommended size of 2000 was considered.

Number of replications. Given dimensional structures (2), sample sizes (2) and ability distributions (5), there were twenty combinations of simulation conditions. 100 test response sets were generated for each combination.

Comparison. Before conducting linking, item parameter estimates were calibrated. Following this, estimates of item characteristics of 20 items were transformed to the initial item parameters. While item estimates of one test are linked to other estimates in practice, item parameters were used as base estimates for evaluation purposes in the present simulation. Three linking methods, i.e., the TCF, Trace, and DDM methods were compared, based on how closely item estimates were transformed to the item parameters (degree of parameter recovery).

Several computer programs were used in the simulation study. In order to generate ability distributions that were bivariate normal with given means and variances/covariances, GENDAT5 (Thompson, 2003) was used. For multidimensional item calibration, a modified Windows version of NOHARM (Fraser, undated; Thompson, 1996) was used. IPLINK (Lee & Oshima, 1996) and MDEQUATE (Li, 1996) were run to implement the TCF and the Trace methods, respectively. For the expansion of the Trace method with a diagonal dilation matrix, a new linking program was written by MATLAB (MathWorks, Inc, 1995).

Evaluation criteria and statistical tests. In the IRT framework, the most popular evaluation criterion for the metric linking is the size of the differences between base estimates and transformed values. Adopting the statistical concepts of accuracy and stability, two summary statistics were used as evaluation criteria: (a) how far transformed values depart from initial item parameters (linking bias), and (b) how much differences fluctuate (root mean square error, RMSE) among items. Linking bias and RMSE were computed by

$$\sum_{i=1}^{n}\frac{(\hat{a}_{ik}^{*}-a_{ik})}{n} \tag{23}$$

$$\left(\frac{\sum_{i=1}^{n}(\hat{a}_{ik}^{*}-a_{ik})^{2}}{n-1}\right)^{1/2}, \tag{24}$$

where, a_{ik} is the ith item parameter on the kth dimension, $\hat{a}_{ik}^{*}$ is the transformed value, and n is the number of items, twenty items for the present simulation. As each item has three parameters (two discriminations and one difficulty-related parameter) and transformed values, there were three sets of biases and RMSEs for each replication.

Because three linking methods were applied to the same simulated response patterns (i.e., three transformation results for each item parameter would be correlated rather than independent), a repeated measures analysis of variance (ANOVA) model was used to detect the effects of simulation conditions (between-factors) and linking methods (within-factor) on bias and RMSE. The model for the bias of first discrimination estimates is

$$Bias(a_1)_{lings} = \mu + \beta_n + \gamma_g + \lambda_s + \gamma\lambda_{gs} + \pi_{i(ngs)} + \alpha_l + \alpha\beta_{\ln} + \alpha\gamma_{\lg} + \alpha\lambda_{ls} + \alpha\gamma\lambda_{\lg s} + e_{li(ngs)}, \quad (25)$$

where $Bias(a_1)_{lings}$ is the bias of the first dimensional discrimination for *l*th linking method, *i*th iteration, *n*th sample size, *g*th group and *s*th structure; μ is the overall mean in the population; β_n is the effect of *n*th sample size (1000 and 2000); γ_g is the effect of *g*th distributional group (groups 1 to 5); λ_s is the effect of *s*th dimensional structure (APSS and MS); $\gamma\lambda_{gs}$ is the interaction effect of group and structure; $\pi_{i(ngs)}$ is the effect of *i*th iteration within *n*th sample size, *g*th group and *s*th structure (1 to 100); α_l is the effect of *l*th linking method (three linking methods); $\alpha\beta_{\ln}$ is the interaction effect of linking method and sample size; $\alpha\gamma_{\lg}$ is the interaction effect of linking method and group; $\alpha\lambda_{ls}$ is the interaction effect of linking method and dimensional structure; $\alpha\gamma\lambda_{\lg s}$ is the interaction effect of linking method, group and dimensional structure; and $e_{li(ngs)}$ is the interaction effect of linking method and iteration within *n*th size, *g*th group and *s*th structure.

Results

In the model of Equation 25, there are three between-factors: sample size, distributional shape, and dimensional structure. The interaction term of between-factors (group by structure) was selected, based on initial examination of full model results. There is one within-factor, linking method, and there are several interaction terms for between- by within-factors. Equation 25 is the model for the bias of the first dimensional discrimination. The same model applies to the bias and log transformed RMSE for all three item parameters. Additionally, in order to obtain a more desirable distribution (i.e., normality), a natural logarithm was taken for RMSEs. Inference statistics of this model tested whether simulation conditions and linking methods had statistically significant effects on the bias and RMSE, and then descriptive statistics of two summary statistics were examined in order to provide more detailed patterns of linking errors.

After finding significant multivariate results (i.e., traditional four statistics such as Pillai' Trace, Wilks' Lambda, Hotelling's Trace, and Roy's largest Root) (Rencher, 1995) for the repeated measures model, univariate test results for six dependent variables were computed in Table 3. In each cell, there are three numbers: F value, degrees of freedom, and eta square (proportion of explained variance to overall variance). It should be noted that the degrees of freedom regarding linking methods for the difficulty parameters are different from those of the discrimination parameters in Table 3. The reason for this is that only TCF and the Trace (or DDM) methods were compared for difficulty parameters because the Trace and DDM methods resulted in the exact same transformations of difficulty parameters.

The results of the repeated measures ANOVA showed that the type of linking method had significant effects on the bias and RMSE of three item parameters, and the soundness of linking results depended on the interactions of simulated testing conditions and the three linking methods.

In order to directly compare behaviors of the three MIRT linking methods across simulation conditions, two summary statistics are plotted in Figures 3 to 8. Each data point of lines is the average of linking errors for 100 replications, and the horizontal axis represents the combinations of five distributional shapes and two dimensional structures. For example, APSS1 indicates the ability distribution of group 1 and APSS items.

The biases of the TCF and the DDM methods were relatively small, and transformations were stable across different sample sizes compared with the Trace method (Figures 3 and 4). Figure 5 shows that the difficulty estimates are over-transformed from the Procrustes rotation, while they were under -transformed with the TCF method. Figures 6 and 7 show that the TCF method provided more stable transformations of discriminations than the two Procrustes based methods, but was more sensitive to changes in the sample sizes. Figure 8 indicates that the transformed difficulty estimates of the Trace and the DDM methods were exactly the same, and they were more stable than the TCF method.

It should be noted that Figures 3 and 4 showed more bias in larger samples, especially in Trace methods. It was more apparent when ability distribution was separated from the default distribution, such as was the case in groups 2 to 5. The reason for this might be that there was a degree of confounding among errors of parameter estimation and scale transformation. As sample size increases, parameter estimates tended to be more accurate (i.e., maintaining dimensional structure), in which case orthogonal rotation and constant dilation made transformation less stable. Similar results of orthogonal rotation were found in previous research (Li & Lissitz, 2000; Min & Kim, 2003). However, error variations were reduced as sample size increased (see Figures 6 to 8). Moreover, there were interactions among linking methods, ability distribution, and dimensional structures. This meant that scale transformations were more stable when the ability

Table 3. *Test Statistics (F), Degrees of Freedom (DF) and Effect Sizes (η^2) of Biases and RMSEs from Repeated Measures ANOVA*

Source	Bias, a_1	Bias, a_2	Bias, d	LN RMSE, a_1	LN RMSE, a_2	LN RMSE, d
Between Factor						
β_n Sample Size	F = 153.84**	156.17**	2.87	459.33**	475.07**	314.24**
	DF = (1,1989)	(1,1989)	(1,1989)	(1,1989)	(1,1989)	(1,1989)
	η^2 =.07	.07	.00	. 19	. 19	. 14
γ_g Distributional Group	403.30**	388.60**	.53	319.57**	412.60**	13.65**
	(4,1989)	(4,1989)	(4,1989)	(4,1989)	(4,1989)	(4,1989)
	.45	.44	.00	. 39	. 46	. 03
λ_s Dimensional Structure	880.61**	628.63**	24.47**	19.40**	.01	.30
	(1,1989)	(1,1989)	(1,1989)	(1,1989)	(1,1989)	(1,1989)
	.31	.24	.01	.01	. 00	.00
$\gamma\lambda_{gs}$ Group × Structure	57.07**	49.81**	1.73	4.73*	2.87*	8.61**
	(4,1989)	(4,1989)	(4,1989)	(4,1989)	(4,1989)	(4,1989)
	.10	.09	.00	.01	.01	.02
Within Factor						
α_l Linking Method	434.90**	234.44**	830.39**	1213.09**	1318.11**	2926.16**
	(2,3978)	(2,3978)	(1,1989)	(2,3978)	(2,3978)	(1,1989)
	.18	.11	.30	.38	.40	.60
$\alpha\beta_{\ln}$ Link × Size	136.09**	125.59**	7.96**	157.50**	173.98**	20.80**
	(2,3978)	(2,3978)	(1,1989)	(2,3978)	(2,3978)	(1,1989)
	.06	.06	.00	.07	.08	.01
$\alpha\gamma_{\lg}$ Link × Group	162.12**	165.01**	2.20	196.03**	216.65**	1.32
	(8,3978)	(8,3978)	(4,1989)	(8,3978)	(8,3978)	(4,1989)
	.25	.25	.00	.28	.30	.00
$\alpha\lambda_{ls}$ Link × Structure	307.82**	274.91**	15.60**	46.79**	85.43**	6.61*
	(2,3978)	(2,3978)	(1,1989)	(2,3978)	(2,3978)	(1,1989)
	.13	.12	.01	.02	.04	.00
$\alpha\gamma\lambda_{\lg s}$ Link × Group × Structure	32.73**	32.04**	3.55**	10.87**	11.26**	8.51**
	(8,3978)	(8,3978)	(4,1989)	(8,3978)	(8,3978)	(4,1989)
	.06	.06	.01	.02	.02	.02

** p<.01, * p<.05

distribution was similar to default one (group 1), correlation and different dilation rates made transformations unstable, the mixed structure resulted in less biased transformation, and these patterns were more apparent in the Trace method.

In general, the TCF and the DDM methods provided less biased metric transformations of discrimination estimates compared with the Trace method. The DDM method with the diagonal dilation matrix significantly reduced linking biases compared with the Trace method, and made more stable transformations for difficulty-related parameters than the TCF method.

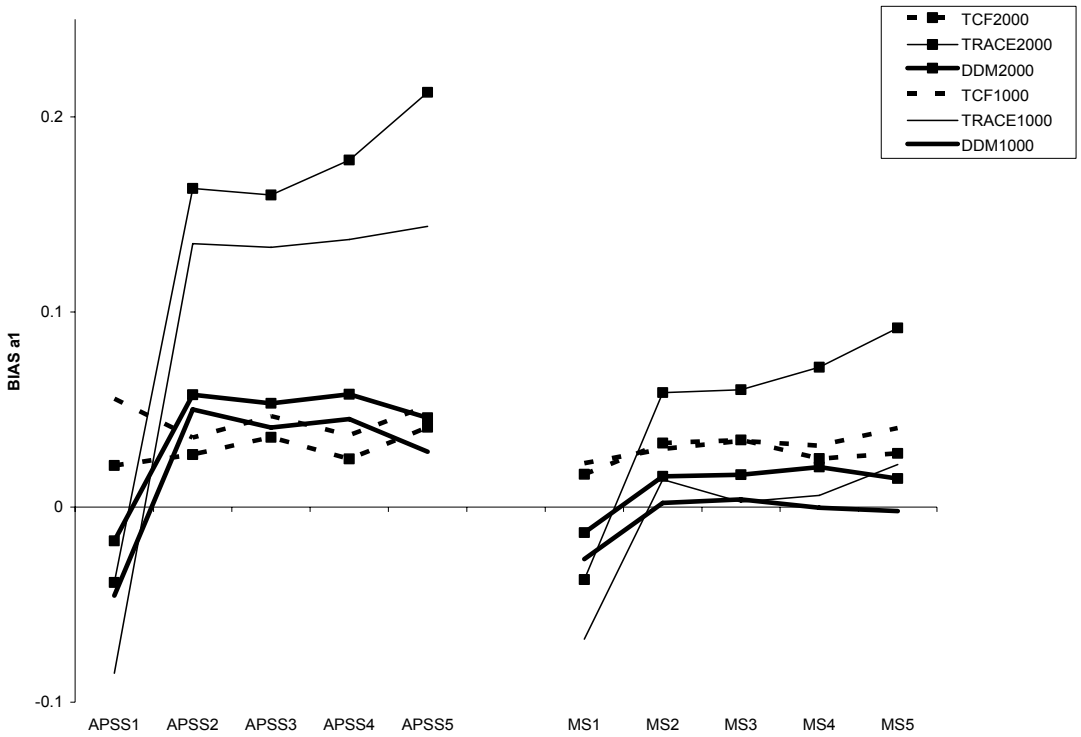


Figure 3. Bias of the First Discrimination. 2000 and 1000 indicate the sample sizes of 2000 and 1000, respectively.

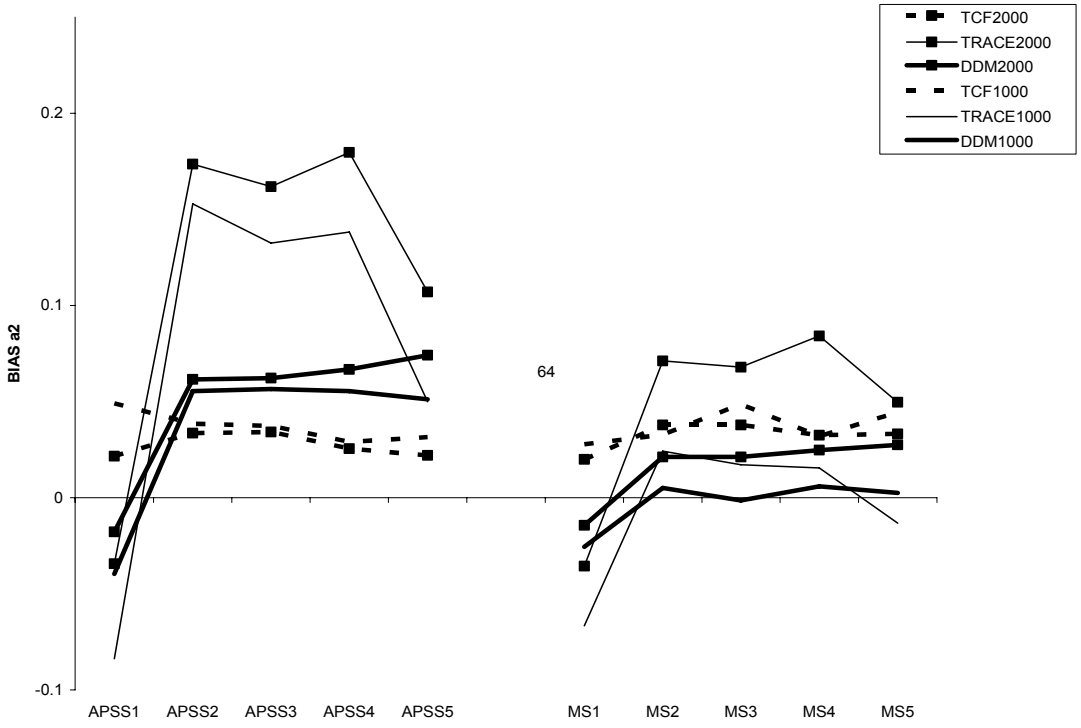


Figure 4. Bias of the Second Discrimination

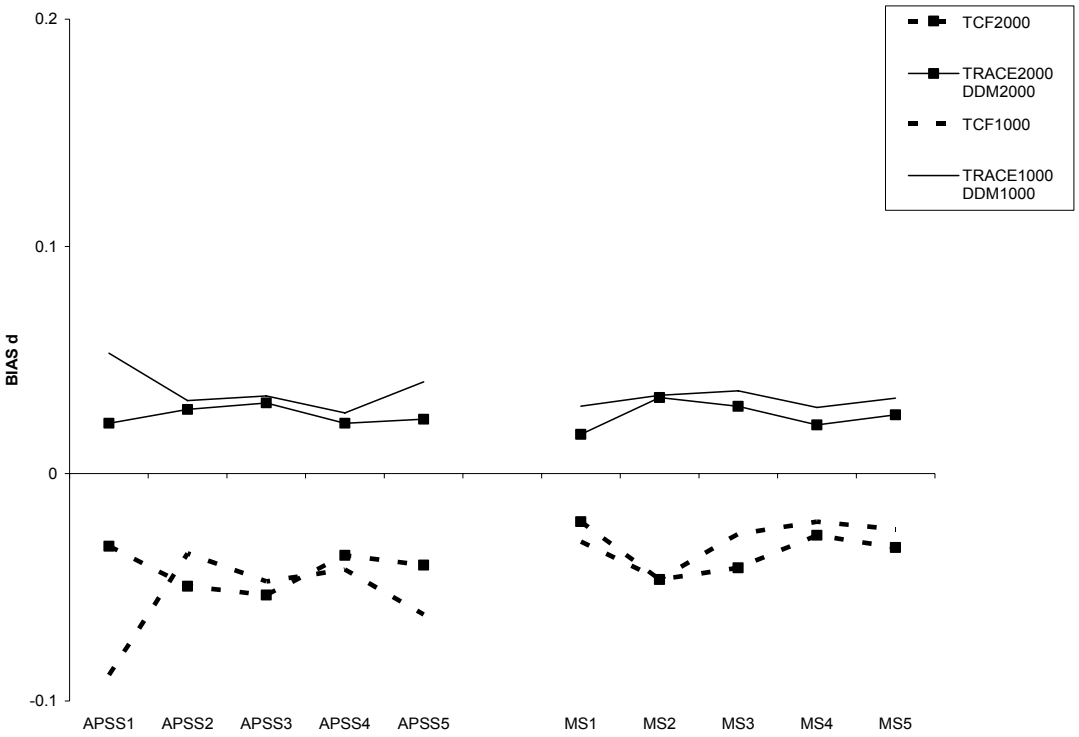


Figure 5. Bias of the Difficulty

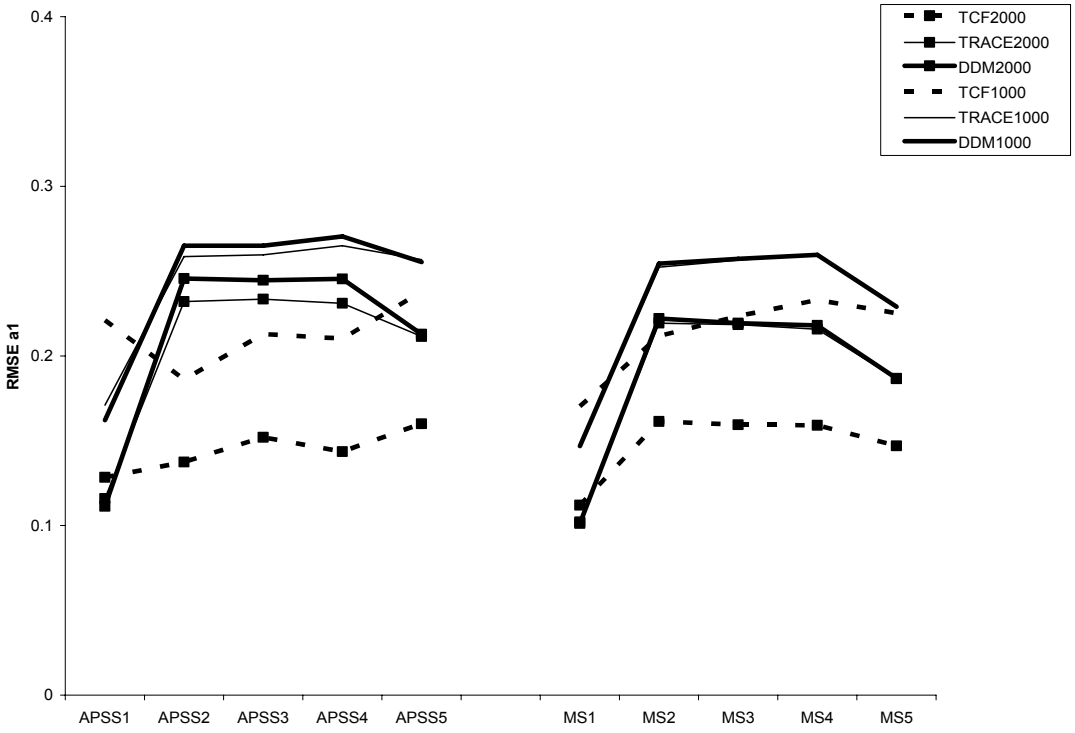


Figure 6. RMSE of the First Discrimination

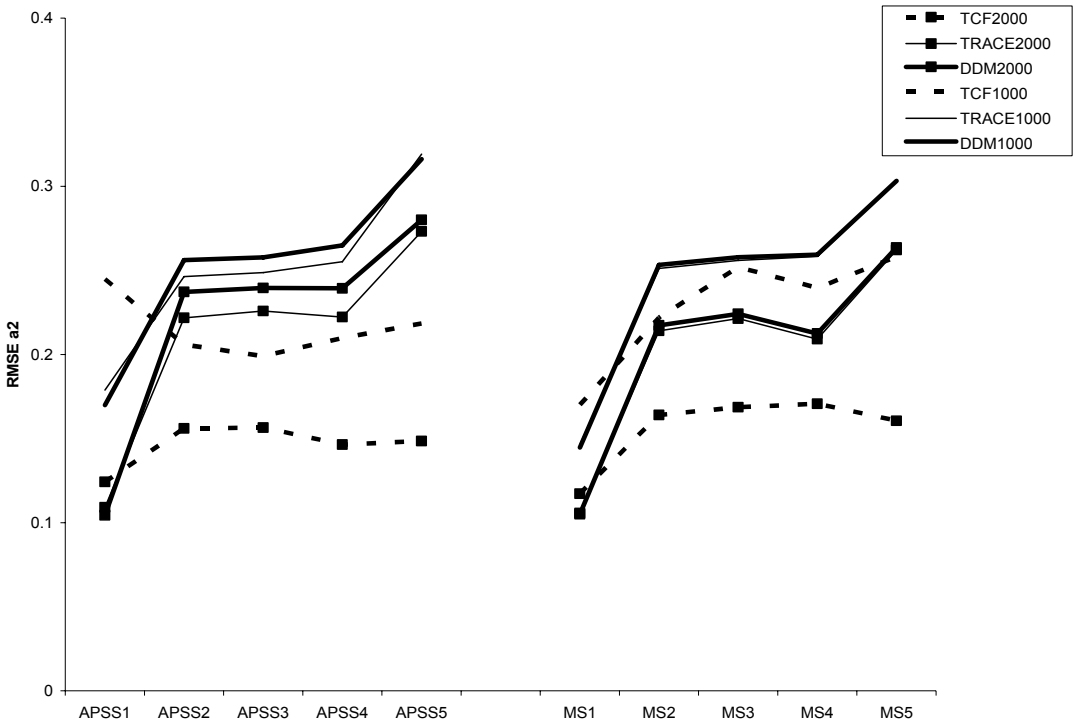


Figure 7. RMSE of the Second Discrimination

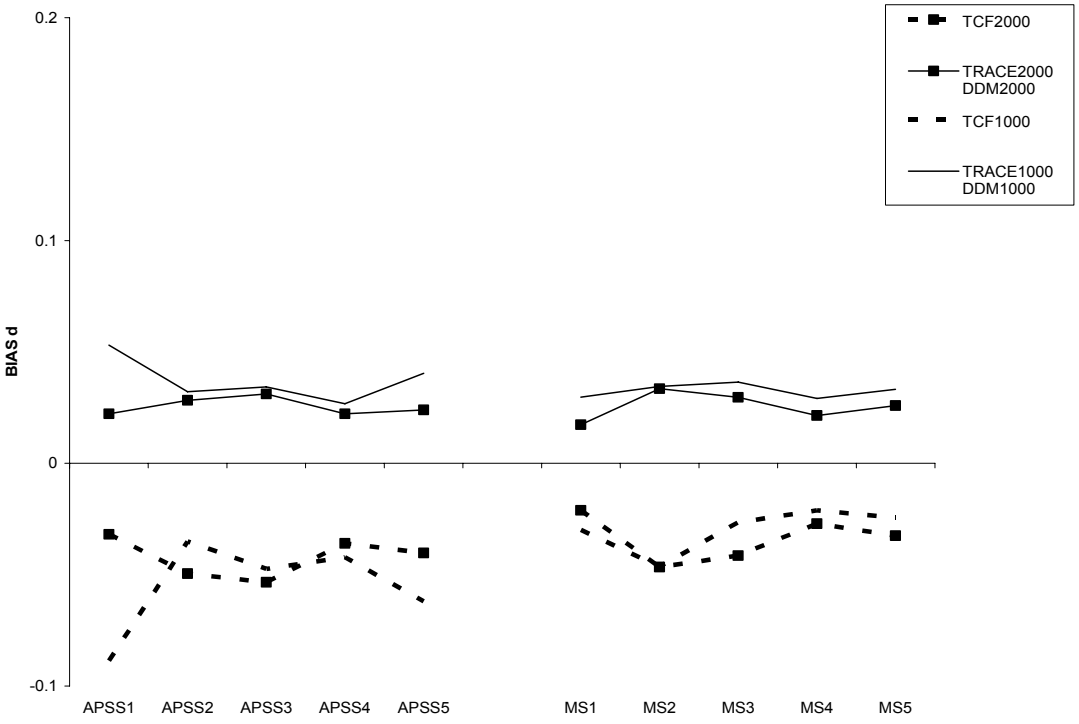


Figure 8. RMSE of the Difficulty

Discussion and Conclusion

The results from this study indicated that modeling unique dilation rates for each ability dimension improved the orthogonal Procrustes metric transformation, which was initially modeled with the Trace method and back to Hirsch's procedures (1989). The TCF and the DDM methods provided more favorable transformation of discriminations than the Trace method, but the two orthogonal Procrustes-based methods produced better difficulty transformations than the TCF method. These differences in the three linking methods can be explained in two ways; types of rotation and estimation criteria.

The rotation matrix of the TCF method adopts the general rotation procedure, i.e., oblique rotation, because it does not put any constraint upon the rotation matrix, while the two Procrustes methods maintain an orthogonal structure. The results of this study showed that the oblique rotation of the TCF method provided closer agreement of dimensional orientations. However, one concern of using an oblique rotation in factor analysis technique, is that the meaning of the reference axes could change after rotation (Harman, 1976). Angles among axes are changed when finding the optimal oblique rotation but the orthogonal rotation maintains the initial structure of a reference system. In the MIRT model context, the orthogonal rotation in the two Procrustes solution-based methods maintain the dimensional structure of test items after conducting metric transformations, while the structure would be somewhat changed with the oblique rotation of the TCF method. However, it is not clear yet whether the item vector structure of the linked test should be maintained through an MIRT metric transformation, or to what degree the oblique rotation of the TCF method really changes the vector structures.

Another distinguishable difference between the two types of methods is the optimization criteria for estimating linking components. Because the TCF method is designed to optimally minimize differences between the linked test response surface and the base one, it outperformed the other methods in obtaining desirable concurrences of true scores. On the other hand, the orthogonal Procrustes-based methods establish an additional problem equation for the translation vector in that these methods were better than the TCF method in transforming the difficulty related parameters. Furthermore, by allowing different dilation rates for different ability dimensions, the DDM method improved linking results compared with the Trace method.

Conclusion The selection of a linking method is a situational specific decision such that it requires personal judgments with knowledge of practical testing conditions and statistical characteristics of linking techniques. In support of this position, the results of this manuscript imply that careful consideration should be made when choosing MIRT linking methods. In addition, a new method of multidimensional scale transformations is proposed which maintains the dimensional structures as well as allowing unique dilation for each dimension. Further research is needed on a variety of levels in order to make MIRT linking methods more practically applicable, such as determining evaluation criteria, separating linking errors from estimation errors, and evaluating the effects of non-normal ability distributions.

References

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20,* 311-329.

Fraser, C. (1988). *NOHARM*: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: University of New England

Harman, H. (1976). *Modern Factor Analysis* (3rd ed.). Chicago: University of Chicago Press.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195-240.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement, 26*, 337-349.

Lee, K., & Oshima, T. C. (1996). *IPLINK*: Multidimensional and unidimensional item parameter equating in item response theory. *Applied Psychological Measurement*, *20*, 230.

Li, Y. H. (1996). *MDEQUATE* [Computer software]. Upper Marlboro MD: Author.

Li, Y. H., and Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT equating. *Applied Psychological Measurement*, *24*, 115 – 138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence.

MathWork, Inc. (1995). *MATLAB: The ultimate computing environment for technical education*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Min, K. S. and Kim, J. P. (2003). A comparison of two linking methods for multidimensional IRT Scale Transformation. *ACT Research Report Series 2003-6*.

Iowa City, Iowa: American College Testing .

Oshima, T. C., Davey, T. C., and Lee, K. (2000). Multidimensional equating: Four practical approaches. *Journal of Educational Measurement*, *37*, 357-373.

Reckase, M. D. (1995). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and Hambleton (Ed.), *Handbook of Modern Item Response Theory*. NY: Springer.

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

Rencher, A. C. (1995). *Methods of Multivariate Analysis.* New York, Wiley.

Roussos, L. A., Stout, W. F., and Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35,* 1-30.

Schönemann, P. H., and Carroll, R. M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrica*, *35,* 245-255.

Stocking, M. L. and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

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.

Thompson, T. (2003). *GENDAT5*: A computer program for generating multidimensional item response data.

Thompson, T., Nering, M., and Davey, T. (1997). *Multidimensional IRT scale linking without common items or common examinees*. Paper presented at the annual meeting of the psychometric society, TN: Gatlinburg,.

Received February 16, 2006
Revision received December 11, 2006
Accepted January 7, 2007