Running Head: LINKING

Observed Score and True Score Equating for Multidimensional Item Response Theory under Nonequivalent Group Anchor Test Design

Ou Zhang

Pearson

M. David Miller

University of Florida

James Algina

University of Florida

Paper presented at the annual meeting of the
American Educational Research Association (AERA) and the
National Council on Measurement in Education (NCME)

April 27-May 1, 2013, San Francisco, CA.

Observed Score and True Score Equating for Multidimensional Item Response Theory under Nonequivalent Group Anchor Test Design

Abstract

For each MIRT ability vector on a particular test form, it is possible that there are an infinite number of ability vectors falling on the equivalent contours of the test characteristic surface on a corresponding equated test form. Therefore, using the number-correct score as the ability measure makes MIRT equating a viable option.

In this study, the equating performances for five MIRT linking methods [i.e., the direct method (OD), the Test Characteristic Function method (TCF), the Item Characteristic Function method (ICF), the Min's method (M), and the non-orthogonal Procrustes method (NOP)] and three MIRT equating procedures [i.e., the full MIRT observed score equating (MOSE), the unidimensional approximation of MIRT true score equating (ATSE), and the unidimensional approximation of MIRT observed score equating (AOSE)] are examined.

Results indicated that the MIRT equating procedures under the TCF, ICF, and OD linking methods showed better equating performance as compared with those under the M or NOP linking methods. The ATSE procedure demonstrated the best performance as compared with the other two equating procedures across all group distribution conditions and all linking methods. The MIRT equating procedures under the NOP linking method demonstrated the worst equating performance within most of the group distribution conditions.

In addition, the group ability mean difference factor had the largest negative effect on the equating results for all three equating procedures across all linking methods.

Observed Score and True Score Equating for Multidimensional Item Response Theory under Nonequivalent Group Anchor Test Design

Introduction

In large scale assessments, multiple test forms with the common-item nonequivalent groups design (NEAT) are widely used to fulfill the test security and fairness requirement. In practice, it is nearly impossible to construct multiple forms that are strictly parallel. So, equating, a statistical process, is used to adjust scores on different test forms so that scores on the forms are comparable (Kolen & Brennan, 2004). Test equating can be categorized as Item Response Theory (IRT) equating or non-IRT equating.

Because the parameter invariance characteristic of IRT offers tremendous flexibility in choosing a plan for calibrating and linking test forms, IRT is widely used in educational measurement.

IRT equating is conducted under the IRT framework. In general, there are three basic steps in IRT equating if the number-correct score is used as an ability measure. These three steps are IRT estimation, IRT linking, and IRT equating (if necessary). IRT estimation is used to estimate the item parameters and ability estimates from different models on the data; IRT Linking is used to transform the parameter scales from different linking methods under the non-equivalent anchor test (NEAT) design and IRT Equating is used to obtain equivalent scores for the different test forms from different equating methods.

Multidimensional item response theory (MIRT) model has been developed in response to the need for modeling the relationship between more than one ability or construct, and also the complexities of the interaction between persons' multiple ability dimensions and items (Reckase,

2005). MIRT models are developed and classified into compensatory and partially compensatory models.

MIRT Linking

Linking or scale aligning is a collection of procedures to put performance or scores on one assessment on a common metric with performance or scores on another assessment. MIRT scale linking is conducted to adjust (1) rotation, (2) correlation, (3) translation (similar to "origin" in UIRT linking), and (4) dilation (similar to "unit of measurement" in UIRT linking). Different methods used different approaches to adjust rotation, translation, and dilation for the parameter estimated coordinate systems so that the scale indeterminacies are taken into account. Similar to unidimensional IRT (UIRT) scale linking, MIRT scale linking is a linear transformation, but the transformation is on multiple dimensions. The MIRT scale linking is graphically displayed below in Figure 1.

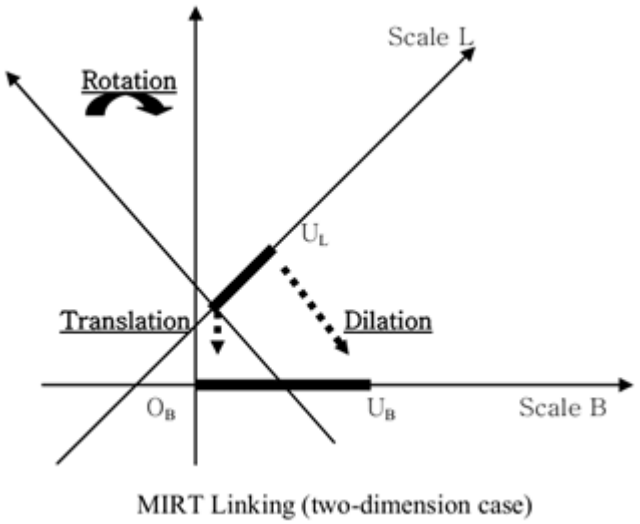

Figure 1. MIRT Linking Components O represents origin, U represents the unit of measurement for Scale L and B. (Adapted from Min, 2003)

Several MIRT scale linking methods have been developed (Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima, Davey, & Lee, 2000; Reckase & Martineau, 2004; Thompson et al., 1997) for the NEAT design, including the Li and Lissitz (2000) method (LL method), the Min (2003) method (M method), the Oshima, Davey, and Lee (2000) method (i.e. Test Characteristic Function or TCF method, Item Characteristic Function or ICF method, Direct Function or OD method), and the Non-orthogonal Procrustes (Reckase & Martineau, 2004) method (NOP method).

These MIRT scale linking methods all use the multiple-dimensional compensatory model. In these methods, three linking coefficients are estimated including a rotation matrix ($\mathbf{A}$) to deal with rotation indeterminacy, a translation vector ($\boldsymbol{\beta}$) and a dilation vector (m) to deal with origin and unit indeterminacy for MIRT scale system (Min, 2003).

These MIRT linking methods differ in: (1) data collection designs; (2) the theoretical foundation to solve rotation indeterminacy (IRT perspective or factor analysis perspective); (3) the rotation approach (orthogonal or non-orthogonal); (3) including or not including the dilation parameters, and; (4) what kinds of dilation parameter the methods have. Furthermore, different methods rely on different mathematical solutions and theoretical perspective to deal with the scale indeterminacies. All current existing MIRT scale linking methods cope with scale indeterminacy by transforming the scale on the rotation, dilation, and translation, either respectively or simultaneously.

The LL method resolves the three indeterminacy problems separately by using a translation vector $\mathbf{m}$, a scalar dilation parameter k, and orthogonal Procrustes rotation matrix $\mathbf{T}$, respectively. The M method improved the LL method by replacing the dilation constant k in the LL method to the diagonal dilation matrix $\mathbf{K}$ that allows for differential dilation/contraction of

the scales of the various dimensions (Min, 2003). The NOP method applies non-orthogonal rotation to correct the weakness in the M method that an infeasible burden of computation exists as dimensionality of test is high.

The TCF method, the ICF method, the OD method, and the NOP method allow using a non-orthogonal rotation approach to solve the rotation indeterminacy problem. In contrast, the LL method and the M method stick to the orthogonal rotation approach.

In the TCF method, the ICF method, the OD method, and the NOP method, the dilation indeterminacy and rotation indeterminacy are solved simultaneously so that no dilation parameters exist in these two methods.

Symmetry Property and Unidimensionalization

The symmetry property of equating, proposed by Lord (1980), requires that the function used to transform a score on the equated form to the base form scale must be the inverse of the function used to transform a score on the base form to the equated form scale (Kolen & Brennan, 2004). However, because ability (i.e., $\hat{\boldsymbol{\theta}}$) is a vector (i.e., $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2 ... \hat{\theta}_m]$) in the MIRT framework, demonstrating equivalence between two ability vectors corresponding to different test forms becomes more complex and is also indirect.

In MIRT, the probabilities of obtaining correct responses to each item are summed to form true scores (i.e., $\tau(\boldsymbol{\theta}) = \sum p(\boldsymbol{\theta})$) in the test characteristic surface (TCS) for each combination of ability levels (corresponding to each dimension). When two test forms are in the same scale metric, the relationship between the location of the ability space and the true score on the test is displayed as the cutoff contour in the TCS. Different ability vectors from two different forms falling on equivalent contours are considered equivalent and may end up with the same true scores.

For a particular true score, an infinite number of combinations of ability levels are associated with that true score. Thus, the ability vectors and their corresponding true scores are no longer symmetric and the symmetry property (Lord, 1980) of equating is violated under MIRT; this can make MIRT equating seems like an impossible task.

One possible solution is to use the number-correct score or true score as the ability measure in MIRT. The process of transforming multidimensional ability vectors into unidimensional measures through a particular MIRT model is a linear combination procedure. This process is called "unidimensionalization" (Zhang, 2012). This linear combination procedure devectorizes the vector or multidimensional features in the MIRT framework. More specifically, when the number-correct score or scale score is used as the ability measure, the MIRT ability vector is unidimensionalized so that the ability measures from different test forms are comparable. As a result of this and most importantly, the symmetry property (Lord, 1980) of equating for two test forms under MIRT is satisfied, and MIRT equating becomes possible.

Unidimensional Approximation

In previous MIRT research (Zhang, 1996; Zhang & Stout, 1999; Zhang & Wang, 1998), researchers claimed that any set of item responses adequately modeled by a multidimensional IRT model could be closely approximated by a unidimensional IRT model with (1) an estimated unidimensional ability composite (Θ_α) and (2) estimated unidimensional item parameters (Zhang & Stout, 1999).

First, a generalized multidimensional compensatory model is defined as:

$$p_{ij}(\boldsymbol{\theta}_i) = H_j(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\theta}_i + d_j) \equiv H_j\left(\sum_{k=1}^{\delta} a_{kj}\theta_k + d_j\right) \quad (1)$$

where "$\mathbf{a}_j^{\mathbf{T}} = (a_{j1}, a_{j2}, ..., a_{jd})$ is the discrimination parameter vector, $a_{j1}, a_{j2}, ..., a_{jd}$ are nonnegative and not all zero, d_j is an index related to the difficulty parameter, and $H_j(x)$ is a link function" (Zhang, 1996, Zhang & Stout, 1999, and Zhang & Wang, 1998).

The ability composite Θ_α of the multidimensional ability vector $\boldsymbol{\Theta}$ (i.e., $\boldsymbol{\Theta} = [\theta_1, \theta_2, ..., \theta_m]$) is defined as a standardized linear combination of $\boldsymbol{\Theta}$. That is:

$$\Theta_\alpha = \hat{\mathbf{a}}^{\mathbf{T}}\hat{\boldsymbol{\theta}} = \boldsymbol{\alpha}^t\boldsymbol{\Theta} = \sum_{j=1}^{d}\alpha_j\theta_j \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_d)^t$ is defined as the direction of composite Θ_α or the unidimensional approximation of the multidimensional ability vector $\boldsymbol{\Theta}$ (i.e., $\boldsymbol{\Theta} = [\theta_1, \theta_2, ..., \theta_m]$), and $Var(\Theta_\alpha) = 1$ is constrained for the scale specification. Additionally, the sum of the direction of composite Θ_α is also defined as 1(i.e., $\sum_{j=1}^{d}\alpha_j = 1$). This approximation is true for any generalized multidimensional compensatory model. Under the assumption that all terms and scores are equally weighted, the formula of the direction of the linear composite can be shown as:

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N}\hat{a}_{jk}}{\sqrt{\sum_{k=1}^{\delta}\left(\sum_{j=1}^{N}\hat{a}_{jk}\right)^2}} \tag{3}$$

where *N* is the total number of items on the test.

Thus, the item parameters of the UIRT approximation for the MIRT model can be obtained as follows:

UIRT approximation discrimination:

$$\hat{a}_{\alpha j} = (1 + \hat{\sigma}_{\alpha j}^2)^{-\frac{1}{2}}\hat{\mathbf{a}}_j^{\mathbf{T}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\alpha}} \tag{4}$$

UIRT approximation index of difficulty:

$$\hat{d}_{\alpha j} = (1+\hat{\sigma}^2_{\alpha j})^{-\frac{1}{2}}\hat{d}_j \tag{5}$$

UIRT approximation item difficulty is obtained with:

$$\hat{b}_{\alpha j} = \frac{-\hat{d}_{\alpha j}}{\hat{a}_{\alpha j}} \tag{6}$$

And variance of the directions for the standardized linear composite Θ_α with:

$$\hat{\sigma}^2_{\alpha j} = \hat{\mathbf{a}}_j^{\mathrm{T}}\boldsymbol{\Sigma}\hat{\mathbf{a}}_j - \left(\hat{\mathbf{a}}_j^{\mathrm{T}}\boldsymbol{\Sigma}\hat{\boldsymbol{\alpha}}\right)^2 \tag{7}$$

The true score of this unidimensional approximation model for the linear composite (θ_α) is defined as T_α. This true score (T_α) associated with the linear composite (Θ_α) is the sum of the probabilities of obtaining correct responses over all items at each composite ability level, and can be mathematically expressed as:

$$T_\alpha = \xi(\theta_\alpha) \equiv \xi\left(\sum_{k=1}^{\delta}\alpha_k\theta_k\right) \tag{8}$$

This expression preserves the properties of unidimensional IRT true scores in that the function $\xi(\cdot)$ is monotonically increasing (e.g., Zhang et al., 1999).

Note that unidimensional approximation is a procedure of unidimensionalization. It is unknown, however, at which step (i.e., IRT estimation, IRT linking, IRT equating) the unidimensionalization should best be conducted. Technically, unidimensionalization is possible at any of the three equating steps. If unidimensionalization was done in the estimation step, a UIRT model would be used to replace the MIRT model so that a set of unidimensional parameters would be obtained for later use in test linking and equating. This procedure is depicted graphically in Figure 2.

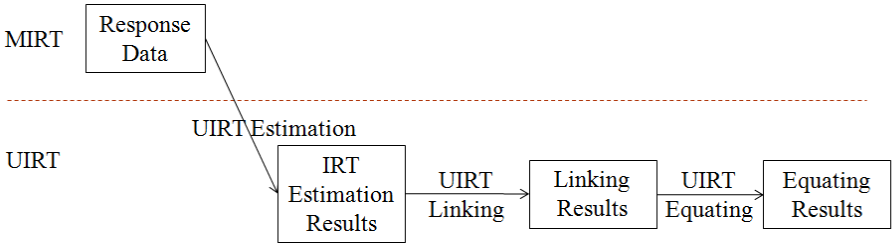

Figure 2. Unidimensionalization at IRT Estimation stage

If the unidimensionalization was conducted in the linking step, the MIRT model would first be estimated and then a unidimensional approximation of MIRT (Zhang, 1996; Zhang & Stout, 1999; Zhang & Wang, 1998) would be conducted so that a UIRT linking and UIRT equating could be applied later. This procedure is depicted graphically in Figure 3.

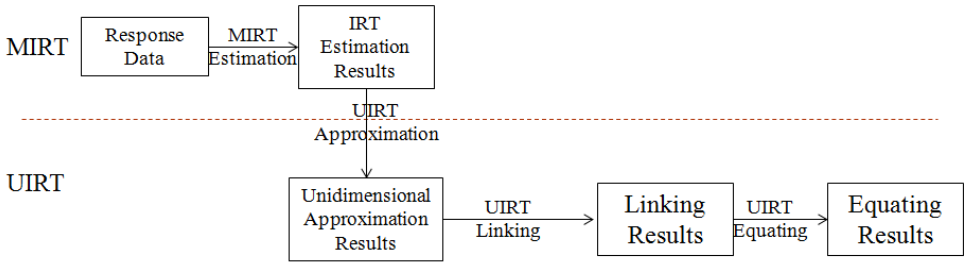

Figure 3. Unidimensionalization before IRT linking

If the unidimensionalization was conducted in the equating step through the compound binomial function from observed score equating method, the multidimensional estimates would be unidimensionalized in the final step. This procedure is depicted graphically in Figure 4.

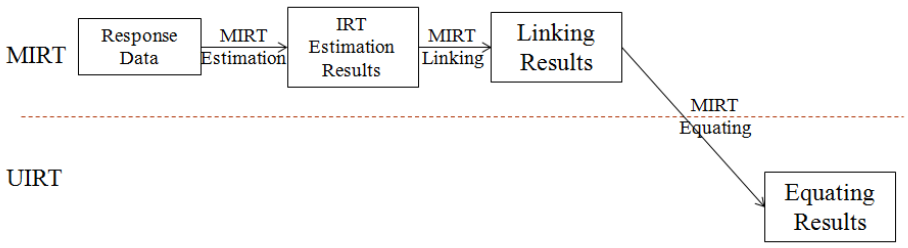

Figure 4. Unidimensionalization at MIRT Equating stage

Finally, the unidimensional approximation (Zhang, 1996; Zhang & Stout, 1999; Zhang & Wang, 1998) could also be used as the unidimensionalization procedure for the equating in the final step; this is depicted graphically in Figure 5.

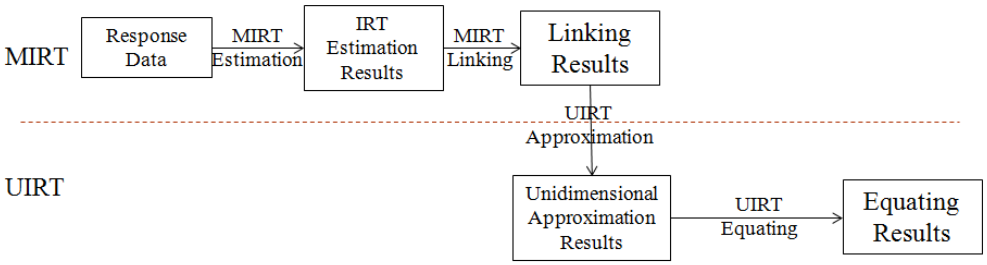

Figure 5. Unidimensionalization before Test Equating stage

MIRT Equating Methods

Several procedures have been developed within the MIRT framework to conduct MIRT equating (Brossman, 2010). These MIRT equating procedures are full MIRT observed score equating (MOSE), unidimensional approximation of MIRT true score equating (ATSE), and unidimensional approximation of MIRT observed score equating (AOSE). The ATSE procedure and the AOSE procedure both apply the unidimensional approximation algorithm (Zhang & Stout, 1999) as their foundation.

Full MIRT Observed Score Equating (MOSE)

The full MIRT observed score equating method (MOSE) is a straightforward extension of UIRT observed score equating. The distribution of observed number-correct scores for examinees of a given ability combination is produced by the compound binomial distribution through a recursion formula (Lord & Wingersky, 1984). The conditional observed score distributions (i.e., $f(x|\boldsymbol{\theta})$) are determined at each combination of ability levels (i.e., the combination of each set of grid points at $\boldsymbol{\theta}$) in the entire ability space (Kolen & Wang, 2007), where $\boldsymbol{\theta}$ is the ability combination vector (i.e., $\boldsymbol{\theta}=[\theta_1, \theta_2, ..., \theta_m]$). In the recursion formula, the

single ability scalar for UIRT is replaced by a vector of the combination of ability levels as follows:

$$\begin{aligned} f_r(x \mid \boldsymbol{\theta}_i) &= f_{r-1}(x \mid \boldsymbol{\theta}_i)(1-p_{ir}) && x = 0 \\ &= f_{r-1}(x \mid \boldsymbol{\theta}_i)(1-p_{ir}) + f_{r-1}(x \mid \boldsymbol{\theta}_i)p_{ir} && 0 < x < r \\ &= f_{r-1}(x \mid \boldsymbol{\theta}_i)p_{ir} && x = r \end{aligned} \tag{9}$$

Next, the conditional observed score distributions (i.e., $f_r(x \mid \boldsymbol{\theta}_i)$) are multiplied by the ability density ($\psi(\theta)$) so that joint distributions of the observed scores are obtained. Once the conditional observed score distributions are determined, they are then multiplied by the multivariate ability density ($\psi(\boldsymbol{\theta})$) to obtain joint distributions of observed scores for the test forms. Finally, the observed marginal distribution ($f(x)$) is determined for each form by either multivariate-accumulation or multiple-integration over all joint distributions at each level of ability combination on the ability space (i.e., $\boldsymbol{\theta}$). The mathematical expression is displayed as:

$$f(x) = \sum_1 \sum_2 \ldots \sum_m f(x \mid \boldsymbol{\theta})\psi(\boldsymbol{\theta}) \tag{10}$$

or

$$f(x) = \int_1 \int_2 \ldots \int_m f(x \mid \boldsymbol{\theta})\psi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{11}$$

where m is defined as the number of dimensions. After these transformations, the traditional equipercentile method is applied to equate both test forms.

Note that the multivariate-accumulation or multiple-integration for obtaining the marginal distribution of observed scores in the final step provides the undimensionalization for the MIRT equating.

Unidimensional Approximation of MIRT True Score Equating (ATSE)

After the unidimensional approximation, the UIRT true score equating procedure is utilized to equate composite true scores (T_α) on both multidimensional test forms. The true score from the base form $\tau_{\alpha B}(\theta_\alpha)$ and the true score from the equated form $\tau_{\alpha E}(\theta_\alpha)$ are considered to be equivalent for a given $\theta_{\alpha i}$. Thus,

$$irt_B(\tau_{\alpha Bi}) = \tau_B(\tau_{\alpha Ei}{}^{-1}) \tag{12}$$

Throughout the iterative procedure (i.e., Newton-Raphson method), the function of the $\theta_{\alpha i}$, is minimized:

$$func(\theta_{\alpha i}) = \tau_{\alpha A} - \sum_{j:A} p_{ij}(\theta_{\alpha i} \mid a_{\alpha j}, b_{\alpha j}, c_j) \tag{13}$$

Finally, using the IRT definition of true score, the composite true score on the base form $\tau_{\alpha B}(\theta_\alpha)$ associated with the composite true score on the equated form $\tau_{\alpha E}(\theta_\alpha)$ can be computed as:

$$\tau_{\alpha B} = \sum_{j:B} p_{ij}(\theta_{\alpha i} \mid a_{\alpha j}, b_{\alpha j}, c_j) \tag{14}$$

Unidimensional Approximation of MIRT Observed Score Equating (AOSE)

The procedure for the unidimensional approximation of MIRT observed score equating is the same as for UIRT observed score equating. After the unidimensional approximation, the conditional distributions for the unidimensional ability composite $f(x \mid \theta_\alpha)$ is determined at each composite ability level (θ_α) through the compound binomial recursion formula (Lord & Wingersky, 1984). Then, the marginal distribution for each observed score is computed by summing or integrating the product of each form's conditional distribution multiplied by the

estimated unidimensional ability distribution in the population of examinees across the estimated unidimensional ability space, shown as:

$$f(x) = \sum_{\theta_\alpha} f(x \mid \theta_{\alpha i}) \psi(\theta_{\alpha i}) \tag{15}$$

or

$$f(x) = \int_{\theta_\alpha} f(x \mid \theta_{\alpha i}) \psi(\theta_{\alpha i}) d\theta_\alpha \tag{16}$$

Finally, the conventional equipercentile procedure is applied to equate the test scores for both forms.

Purpose

Although three MIRT equating procedures were recently developed by Brossman (2010), no simulation studies have been conducted to determine how these procedures perform under a variety of settings. Therefore, the performance of MIRT equating procedures under the NEAT design requires further investigation. In this study, we examine the performance of MIRT equating procedures under the NEAT design and explore how different MIRT linking methods interact with these equating procedures to impact equating results under various testing conditions. Five MIRT linking methods (i.e., the direct method, the Test Characteristic Function method or TCF, the Item Characteristic Function method or ICF, the Min's method or M, and the non-orthogonal Procrustes method or NOP) and three MIRT equating procedures (i.e., MOSE, ATSE, and AOSE) are examined.

Method

For this study, the compensatory two-parameter two-dimensional logistical model (M2PL) was selected as the MIRT model. The computer program TESTFACT (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, 2003) was used for MIRT estimation.

Linking and equating procedures were conducted through the statistical language R 2.14 (R Development Core Team, 2007). Five MIRT linking methods were investigated: the OD method, TCF method, and ICF method from Oshima et al. (2000); the M method from Min (2003); and the NOP method from Reckase and Martineau (2004). Three MIRT equating procedures proposed by Brossman (2010) were also examined: the MOSE method, the ATSE method, and the AOSE method.

Data

The item response data were generated using the statistical language R 2.14 (R Development Core Team, 2007). The design used for data generation was a two-factor, completely crossed design with 4 (ability distributions) $\times$ 2 (test structures) for a total of 8 data generation conditions. Response data were replicated 200 times from a set of population item parameters (40 items) and the sample size for each group were set to be equal to 2,000 for each condition.

In all conditions, the ability dimensions were uncorrelated in the base group. Four conditions were created by varying the means, variances, and correlations between the ability dimensions for the equated groups: (1) No difference in the base and scaled groups (the null condition), (2) differences in θ variances, (3) differences in θ means, and (4) differences in θ correlation. The details of the population design are shown in Appendix Table A-1.

The length of total test was set as 40, and 20 of those items (i.e., 50% of the total items) were used as the common/anchor test section. Approximate simple structure (APSS) and complex structure (CS) were applied in this study, and the approximate simple structure and complex structure item parameters for the base form unique item section, the equated form

unique item section, and the anchor item section are presented in Appendix Table A-2, Table A-3, Table A-4, Table A-5, and Table A-6, respectively.

Synthetic Population Weights and Criterion Equating Method

Synthetic population weights for population 1 and 2 were used to define the target population, and were set as $w_1 = .5$ and $w_2 = .5$. A very large sample (e.g., 200,000) was treated as a population (Harris & Crouse, 1993), and sample groups of examinees were drawn from that population and used to evaluate the different equating methods by comparing the results.

The frequency estimation method for the NEAT design was used as a criterion equating function for comparing the MIRT equating procedures, since this method only employs total test scores and the assumptions associated with this procedure were not expected to be violated in this study.

Evaluation Criteria

According to previous equating literature (e.g., Harris & Crouse, 1993; Zeng & Kolen, 1995), the evaluation criteria for equating results include: Standard Error of Equating conditional on scores (*SEE*), equating bias (Livingston, 1993), and Root Mean Square Deviation (*RMSD*) for each score point. Weighted average bias ($Bias_w$) and weighted average Root Mean Square Deviation ($ARMSD_w$) for the entire test form were used as criteria to evaluate the equating methods in this study. Due to limitations of space, only $Bias_w$ and $ARMSD_w$ were the only evaluation criteria reported for this study.

Equating bias

Equating bias is defined as the mean difference between an equating method and criterion equating function (i.e., Frequency Estimation method) over N replications. The bias at each raw score point x_i is defined as:

$$Bias_i = \frac{\sum_{k=1}^{N}\left[\hat{e}_{base_k}(x_i) - e_{base}(x_i)\right]}{N} \tag{17}$$

where $e_{base}(x_i)$ indicates the raw score equivalent calculated from the criterion equating function.

Root mean square deviation ($RMSD$)

The Root Mean Square Deviation ($RMSD$) is a measure of the overall equating accuracy. It is defined as:

$$RMSD_i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[\hat{e}_{base_k}(x_i) - e_{base}(x_i)\right]^2} \tag{18}$$

where $\hat{e}_{base_k}(x)$ denotes the raw score equivalent calculated from one equating procedure in replication k and $e_{base}(x_i)$ indicates the raw score equivalent calculated from the criterion equating function.

Weighted average equating bias ($Bias_w$)

Weighted Average Equating Bias ($Bias_w$) is used to evaluate the systematic error in equating for each equating procedure as compared to the criterion equating function at the test level. The weighted average equating bias ($Bias_w$) for over all available score points was computed as:

$$Bias_w = \sum_{x=1}^{39} Bias\left[\hat{e}_{base}(x_i)\right]P(x_i) \tag{19}$$

Weighted average root mean square deviation ($ARMSD_w$)

Weighted Average Root Mean Square Deviation ($ARMSD_w$) is used to evaluate the discrepancy between each equating procedure and the criterion equating function at the test level.

The weighted average root mean square deviation ($ARMSD_w$) for over all available score points was computed as:

$$ARMSD_w = \sum_{x=1}^{39} RMSD[\hat{e}_{base}(x_i)]P(x_i) \quad (20)$$

where $P(x_i)$ is the proportion of examinees from the target population who have an observed score of x_i on the equated form.

ANOVA Analysis

Because five MIRT linking methods and three MIRT equating methods were applied to the same response patterns, a repeated ANOVA model was used to detect the effects of simulation conditions (between-factors), linking methods (within-factors), and equating methods (within-factors) on the weighted average root mean square deviation ($ARMSD_w$) and weighted average bias ($Bias_w$) for each iteration.

Two summary statistics were examined to provide detailed patterns of errors associated with MIRT linking, equating, group distribution difference and test structure. The proportion of variance effect size of partial ω^2 was reported and interpreted in this study.

Results

In the first section, the summary of ANOVA analysis results (i.e., ω^2) for weighted root mean square deviation ($ARMSD_w$) and weighted bias ($Bias_w$) for the entire test are presented. In the second section, the results of comparisons for linking method and group distribution interaction, and the results of comparisons for equating method and group distribution interaction are presented.

The proportion of variance effect size of ω^2

The results of ANOVA tests for all four factors are presented in Table A-7. The proportion of variance effect sizes of ω^2 for both weighted average bias and weighted average RMSD indicate that the effects of both linking and equating results were most dependent upon group distribution differences. The largest effect size was the interaction of linking method with group distribution factor, with an effect size of partial ω^2 equal to .88045 for $Bias_w$ and .94122 for $ARMSD_w$, respectively. The second largest effect size was the interaction of equating method with group distribution factor, with an effect size of partial ω^2 equal to .46236 for $Bias_w$ and .58711 for $ARMSD_w$, respectively. Test structure and all the interactions that include test structure had a very small effect size of total ω^2.

This pattern is made clearer by the results of effect size of partial ω^2 presented in Table A-7. That is, the linking method × group distribution interaction accounted for the largest portions of $Bias_w$ and $ARMSD_w$ in the equating results for the entire test. Also, the equating method × group distribution interaction accounted for the second largest portions of $Bias_w$ and $ARMSD_w$ in the equating results for the entire test.

In sum, results of the repeated measures ANOVA showed that group distribution differences and type of MIRT linking had significant effects on equating results (i.e., linking method × group distribution interaction). Test structure and all the interactions including test structure had very small effects on equating results. And the soundness of equating results depended on various group distribution differences, linking methods, equating methods, and their interactions.

Comparison for the Linking Method x Group Distribution Interaction

Since the linking method × group distribution interaction accounted for the largest portions of $Bias_w$ and $ARMSD_w$ in equating results for the entire test, the mean of $Bias_w$ and $ARMSD_w$ for the equating results were obtained by averaging the mean results from different test structures and MIRT equating methods; this was done in order to directly compare the equating results of five MIRT linking methods across different group distribution conditions. These results are shown in Table A-8, Table A-9 and Figure 6, Figure 7.

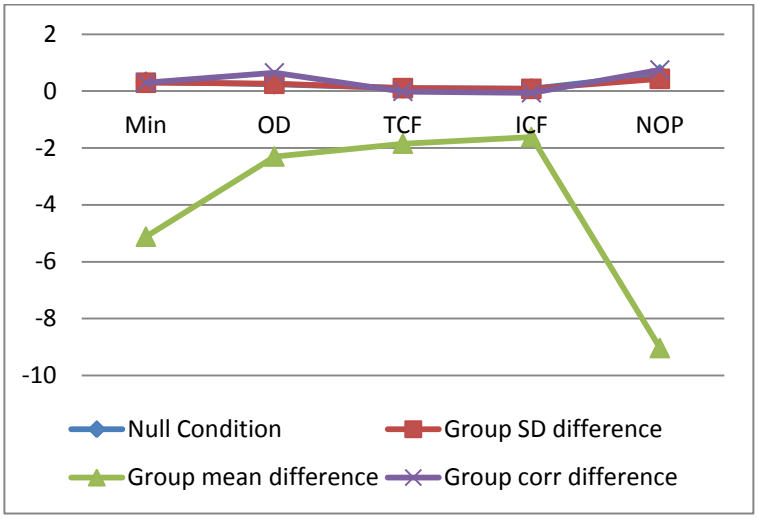

Figure 6. Weighted mean Bias for linking methods × group

In general, the TCF method and the ICF method performed best across all group distribution conditions with means equal to -.4155, -.3799 for $Bias_w$ and 2.0080, 1.3811 for $ARMSD_w$, respectively, as compared with the other three linking methods. The OD method and the M method had less biased and more stable results than the NOP method in terms of smaller means of $Bias_w$ and $ARMSD_w$ on equating results; more specifically, means were equal to -.2896, -1.0497 for $Bias_w$, and 2.4079, 7.4168 for $ARMSD_w$, respectively, for OD and M methods. The NOP method performed worst of all and had the largest means of $Bias_w$ and $ARMSD_w$ (-1.8192

and 22.6127), as compared with the other four linking methods. That is, under the NOP method the equating results for the entire test had the largest amount of $ARMSD_w$ (22.6127), which was approximately the same as the total test scale. More specific results follow.

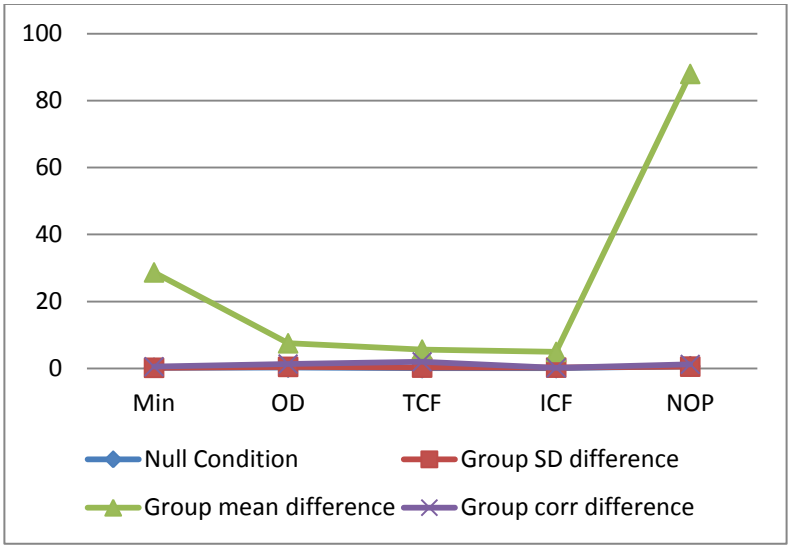

Figure 7. Weighted mean ARMSD for linking methods × group

Under the condition with only group differences in standard deviation (i.e., Group 2), the performance pattern across all five linking methods was similar to the general pattern, but the M method performed best as compared with the other four linking methods, especially in terms of smallest magnitude of mean of $ARMSD_w$; means were equal to .29248 for $Bias_w$ and .20884 for $ARMSD_w$.

Under the condition with group mean differences (i.e., Group 3), the magnitude of means of $Bias_w$ and $ARMSD_w$ for all five MIRT linking methods drastically increased as compared with Group 1 and Group 2. The results of $Bias_w$ for the M method, the OD method, the TCF method, the ICF method, and the NOP method under condition 3 were 5.119, -2.3007, -1.8481, -1.6189, and -9.0351, respectively. The results of $ARMSD_w$ for the M method, the OD method, the TCF method, the ICF method, and the NOP method under condition 3 were 28.6865, 7.52435,

5.64556, 4.9614, and 87.9546, respectively.The performance pattern across all five linking methods was similar to the general pattern but had much larger magnitude of means for $Bias_w$ and $ARMSD_w$. Under the NOP method, the average results of $ARMSD_w$ was 87.9546, which means the discrepancy between the average scores of the three MIRT equating procedures and the criterion equating function scores was more than double that of the entire test scale.

Under the condition where correlation exists between group ability dimensions (i.e., Group 4), the ICF method outperformed the other four linking methods in terms of having the smallest mean of $ARMSD_w$ (.19166). Generally, the results of $Bias_w$ and $ARMSD_w$ for equating results with all five linking methods under condition 4 were comparatively small, with the means equal to .321992 for $Bias_w$ and 1.06243 for $ARMSD_w$. The performance pattern of all five linking methods under correlated ability dimensions was similar to the general pattern across all group distribution conditions.

Comparison for the Equating Method x Group Distribution Interaction

After the linking method $\times$ group distribution interaction, the equating method $\times$ group distribution interaction accounted for the next largest portions of $Bias_w$ and $ARMSD_w$ in the equating results for the entire test. The means of $Bias_w$ and $ARMSD_w$ for the equating results were obtained by averaging the mean results from different test structures and MIRT linking methods, for the purpose of directly comparing equating results obtained with the three MIRT equating methods across different group distribution conditions. These results are shown in Table A-10, Table A-11 and Figure 8, Figure 9.

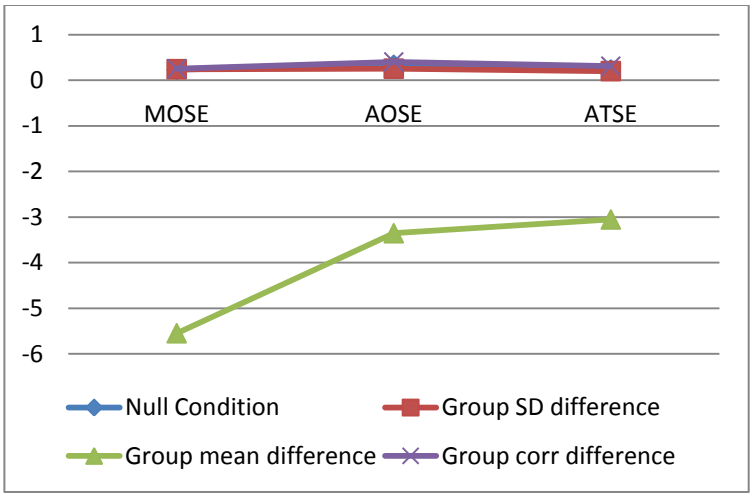

Figure 8. Weighted mean Bias for equating methods × group

Overall, it was found that all three MIRT equating methods performed comparatively well when there was no group mean difference, including under the null condition. Generally, the ATSE method demonstrated the best equating performance, with means equal to -.5774 for $Bias_w$ and 5.0167 for $ARMSD_w$, as compared with the other two equating methods (i.e., MOSE, AOSE) across all group distribution conditions. The MOSE method displayed the worst equating performance, with means equal to -1.2047 for $Bias_w$ and 10.6396 for $ARMSD_w$.

With no group distribution mean differences, the MOSE method performed better than the AOSE method in terms of smaller values of the mean of $Bias_w$ and $ARMSD_w$ for the equating results for the entire test (i.e., .23211 for $Bias_w$ and .23195 for $ARMSD_w$ under the null condition). When there was a group distribution mean difference, both the ATSE method and the AOSE method outperformed the MOSE method with regard to smaller values of the means of $Bias_w$ and $ARMSD_w$. However, the magnitude of the means for $Bias_w$ and $ARMSD_w$ drastically increased for all three MIRT equating methods when group mean differences existed. For example, under group distribution condition 3, the mean results of all three MIRT equating

procedures were -2.23063 for $Bias_w$ and 16.3327 for $ARMSD_w$, as compared with .36066 for $Bias_w$ and .372772 for $ARMSD_w$ under the null condition. More specific results follow.

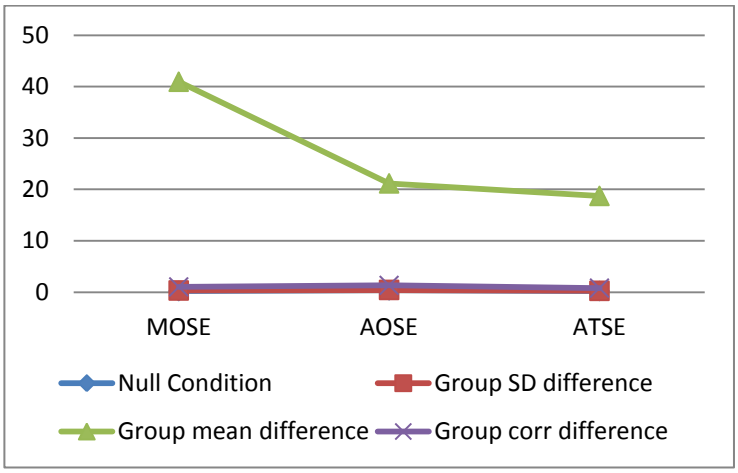

Figure 9. Weighted mean ARMSD for equating methods × group

When only the group standard deviation varied (i.e., Group 2) or only a correlation existed between group ability dimensions (i.e., Group 4), the performance pattern across all three equating methods was similar to the general pattern with means equal to -2.23063 for $Bias_w$ and 16.3327 for $ARMSD_w$. The ATSE procedure outperformed the other two equating methods with the smallest magnitude of mean for $ARMSD_w$ (i.e., .27605).

Under the condition with group mean differences (i.e., Group 3), the values of the means of $Bias_w$ and $ARMSD_w$ for all three MIRT equating methods greatly increased, with the means equal to -2.2306 for $Bias_w$ and 16.3327 for $ARMSD_w$, as compared with means of .2989 for $Bias_w$ and .37375 for $ARMSD_w$ under conditions without group mean differences (i.e., Group 2). The performance pattern across all three equating methods was similar to the general pattern but with larger magnitude of means for $Bias_w$ and $ARMSD_w$.

Discussion and Conclusion

Equivalent Score Difference

In this study, MIRT equating under the NEAT design was compared under eight conditions with four different group distributions across two different test structures. From these results, some characteristics of MIRT equating with the NEAT design can be identified.

First of all, test structure and all the interactions including test structure had a very small effect on equating results. Second, among all three group distribution factors (i.e., group mean, correlation and standard deviation), the group mean factor influenced equating results the most. The group correlation factor and the group standard deviation factor had a similar level of effect on the equating results, but their impact was not as large as the group mean factor. Third, the interaction of group distribution differences and type of MIRT linking method had a huge effect on the equating results. Fourth, the interaction of group distribution differences and type of MIRT equating procedure also had a large effect on the equating results.

All three MIRT equating procedures performed best under the TCF and the ICF MIRT linking methods when there were significant group distribution differences. When group distribution differences existed, equating results had smaller discrepancies under the OD and the M methods than they did under the NOP method. The equating procedures under the NOP method had the lowest robustness when there were group distribution shape differences. This was consistent with results found in previous studies (Simon, 2008). Moreover, MIRT equating procedures demonstrated the worst performance under the NOP linking method with larger score differences across score scale.

In this study, some interesting results were found by comparing MIRT equating procedures with the criterion equating function within each population condition through equivalent score differences. Results of the comparison among the three MIRT equating procedures were

obtained by averaging equating results from each MIRT equating procedure across all linking methods.

First, the ATSE procedure demonstrated, overall, the best equating performance as compared with the other two equating procedures (i.e., MOSE and AOSE) across all group distribution conditions. Second, all three MIRT equating procedures performed comparatively well when no group mean difference existed, especially under the null condition. Third, the MOSE procedure performed better than the AOSE procedure in terms of the equivalent score difference across score scale when no group distribution mean differences existed. Fourth, both the ATSE procedure and the AOSE procedure outperformed the MOSE procedure when there were group distribution mean differences. However, when group mean differences existed, the equating results for all three MIRT equating procedures had larger discrepancies than those under conditions with no group mean differences.

Fifth, the ATSE procedure performed better than the other two equating procedures when only the group standard deviation varied or only a correlation existed between group ability dimensions. When group means differed, the discrepancies between equivalent scores from all three MIRT equating procedures and equivalent scores from the criterion equating function greatly increased. But the ATSE procedure also outperformed the other two equating procedures in terms of smaller equivalent score differences.

Because MIRT observed score equating and MIRT true score equating are defined differently, the observed score and the true score equating procedures are not expected to perform similarly even under ideal conditions. When two groups are non-equivalent, the ATSE procedure had, overall, the best equating performance as compared with frequency estimation equating results in this study. This was true even though the frequency estimation equating

procedure is an observed score procedure. It is currently unknown as to why the ATSE procedure performed best among all three MIRT equating procedures, even when compared with the observed score criterion equating function. This result might have been caused by group ability non-equivalence. More specifically, since the MIRT true score procedure is sample invariant and MIRT observed score procedures may be influenced by sample variation, it is possible that the MIRT true score procedure outperforms the MIRT observed score equating procedures as group non-equivalence exists, such as with the NEAT design.

Possible Effects of Equivalent Score Difference

Effects from IRT Estimation

The first possible effect that could influence MIRT equating results is the MIRT estimation process itself. In this study, item calibration was done by using TESTFACT to obtain MIRT parameter estimates. TESTFACT provides two types of rotation solutions in its IRT estimation process: the orthogonal 'VARIMAX' rotation solution and the non-orthogonal 'PROMAX' rotation solution. According to previous literature (i.e., Li & Lissitz, 2000), the 'PROMAX' rotation solution is recommended because the calibrated item parameter estimates can be rotated obliquely for better interpretation. For that reason, the 'PROMAX' solution was selected for use in this study.

However, using the 'PROMAX' rotation solution in the MIRT estimation process might be controversial. On one hand, using 'PROMAX' may provide better-interpreted item parameter estimates (Li & Lissitz, 2000); on the other hand, under such oblique rotation the overall discrimination power for each item (which is related to the geometric length of the item as represented in multidimensional space) may change, since each item is rotated obliquely. Furthermore, the MIRT difficulty parameter (d) may vary accordingly. That is, it is possible that the direction of best measurement for the entire test may change by using the 'PROMAX'

rotation solution. Thus, using the 'PROMAX' rotation solution in the MIRT estimation process may affect the MIRT linking process.

Since most constructs (and dimensions within a construct) are correlated in education and psychology, correlated ability dimension conditions are included in this study. Because of the characteristics of the IRT estimation in TESTFACT, correlations among item scores are accounted for solely by the a parameters (Li, 1997; Reckase, 1997; Wei, 2008). However, by using the 'PROMAX' rotation solution in the MIRT estimation process, correlations among the item scores may not be solely accounted for by the a parameters. This may also affect the MIRT linking process. It is currently unknown which rotation solution will have a greater effect on MIRT equating procedure performance. Furthermore, the amount of error due to item parameter estimates from the MIRT estimation was not examined separately from equating errors in this study. Thus, further investigation into how MIRT estimation may affect equating results is warranted.

Effects from IRT Linking Methods

The process of MIRT linking may also affect MIRT equating results. As mentioned previously, different linking methods apply different types of rotation, dilation, and translation approaches. Also, different linking methods utilize different types of optimization approaches in the MIRT linking process. Therefore, applying different MIRT linking methods in MIRT equating procedures may result in different equating performance, even within the same MIRT equating procedure.

In the TCF, the ICF, and the OD linking methods, the rotation matrix $\mathbf{A}$ and translation vector $\boldsymbol{\beta}$ are optimized simultaneously through the linking process. In the M linking method, only the rotation matrix $\mathbf{A}$ is optimized by minimizing the trace function for the product of the least

square difference ($\mathbf{E}_1$) and its transpose ($\mathbf{E}_1'$). In the NOP linking method, no optimization process is applied and the MIRT linking process is done solely through the non-orthogonal Procrustes procedure. Thus, the number of optimization processes included in the linking methods may be an underlying reason for differences in equating performance. Accordingly, it is possible that because both the rotation matrix $\mathbf{A}$ and translation vector $\boldsymbol{\beta}$ are optimized in the MIRT linking processes, that explains why the MIRT equating procedures under the TCF, the ICF, and the OD linking methods demonstrate better equating performance as compared with the M and NOP linking methods. It is also possible that because no optimization process is applied in the NOP process, the MIRT equating procedures under the NOP linking method demonstrate the worst equating performance among all MIRT linking methods. In this study, the amount of error due to MIRT linking was not examined separately from equating errors. Therefore, further investigation into the impact that different MIRT linking methods may have on equating results is suggested.

Limitation and Future Research Direction

This study is the first simulation study to evaluate the performance of different MIRT equating procedures. Specifically, this study explores the performance of multiple MIRT equating procedures under the NEAT design. It should be noted that we could have considered more comprehensive factors. For example, more sophisticated combinations of different populations could have been included. Since it was impossible to include all these factors, only a few of them were considered; thus, this study is limited by the restricted number of conditions considered.

Also, the IRT software used in this study was TESTFACT. As mentioned previously, TESTFACT only provides limited options for rotation solutions in its IRT estimation process.

Possible problems may have been created by the selection of rotation type in the process of MIRT item calibration. Future research should consider additional choices for rotation in the MIRT estimation process, from different MIRT software. Moreover, when there are high correlations between ability dimensions and non-equivalent groups, the choice of a rotation solution for MIRT item calibration becomes much more complex. Better estimation procedures with correlated ability dimensions and non-equivalent groups are needed. Thus, the possibility of comparing multiple programs for MIRT (Mplus, BMIRT, and IRTPRO) needs to be considered in future studies.

Mean ability differences between groups had the greatest influence on the equating results for all three equating procedures across all linking methods. This is likely due to the violation of the population invariance requirement for equating. Also, it may have been impacted by the fact that no optimization is involved in the translation in any of the linking methods, such that the adjustment process in MIRT linking may not work effectively.

Another limitation may be that the rotation in the MIRT linking process used in this study is controversial. On one hand, only orthogonal rotation in MIRT linking is recommended in the literature (Brossman, 2010; Min, 2003). In Brossman's study, the author stated that although the discrimination parameters changed through the orthogonal rotation, the overall discrimination power and the MIRT difficulty parameter for each item remained the same. In Min's study (Min, 2003), concerns about using oblique rotations in the MIRT linking process were addressed. He believed that the meaning of the reference axes could change after oblique rotation because the angles among axes are changed when finding the optimal rotation, while the orthogonal rotation maintains the initial structure of a reference system. Neither Brossman nor Min recommend using oblique rotation in MIRT linking process, something that was proposed by Oshima, Davey

and Lee (2000). On the other hand and based on the MIRT results obtained from this study, MIRT equating procedures performed under the oblique rotated linking methods (TCF, ICF, and OD) demonstrated better equating performance than those performed under the orthogonal rotated linking methods (M). It is not clear whether researchers need to maintain item vector structure through an orthogonal rotation, nor is it clear to what extent the oblique rotation used in most of the linking methods changes the vector structures such that the performance of the MIRT equating procedures is influenced. Therefore, further investigation into what types of rotation used in the MIRT linking process for MIRT equating is needed.

Lastly, it is worth noting that although test forms to be equated are typically designed to cover the same content domain, the multidimensional feature of some tests implies that different total scores across the entire score scale might carry different weights from different dimensions for each population. This may be true even though a unidimensionalization procedure is conducted in the process to obtain total scores. Due to limits of time and space, this issue was not discussed in this study. Therefore, further research into this issue in MIRT equating is recommended.

LIST OF REFERENCES

Brossman, B. G. (2010). Observed score and true score equating procedures for multidimensional item response theory. *Unpublished doctoral dissertation,* University of Iowa. http://ir.uiowa.edu/etd/469.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (1999). *TESTFACT 3: Test scoring, items statistics, and full-information item factor analysis.* Chicago: Scientific Software International.

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.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (Second ed.). New York: Springer.

Kolen, M. J., & Wang, T.-Y. (2007*). Conditional standard errors of measurement for composite scores using IRT.* (Unpublished manuscript).

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

Livingston, S. A. (1993). Small-samp le equating with log-linear smoothing. *Journal of Educational Measurement, 30,* 23-39.

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

Lord, F. M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological measurement, 8,* 452-461.

Min, K. S. (2003). The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations. *Unpublished Dissertation,* Michigan State University, East Lansing, MI.

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

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3- 900051-07-0, URL http://www.R-project.org.

Rechase, M. D., & Martineau, J. (2004, October). *The vertical scaling of Science Achievement Tests.* Paper commissioned by the Committee on Test Design for K-12 Science Achievement, Center for Education, National Research Council.

Reckase, M. D. (2005). Multidimensional item response theory models. In Kimberly Kempf-Leonard (Ed.) *Encyclopedia of Social Measurement*, (Vol. 2, pp. 771-777). San Diego, Calif.; London: Academic.

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

Simon, M. K. (2008). Comparison of concurrent and separate multidimensional IRT linking of item parameters. *Unpublished Dissertation,* University of Minnesota.

Thompson, T. D., Nering, M., & Davey, T. (1997). *Multidimensional IRT scale linking without common items or common examinees.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Zeng, L. & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19(3).* 231-240.

Zhang, J. (1996). Some fundamental issues in item response theory with applications. *Unpublished doctoral dissertation,* University of Illinois at Urbana-Champaign, Department of Statistics.

Zhang, J., & Stout, W. F., (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64,* 129-152.

Zhang, J., & Wang, M. (1998, April). *Relating reported scores to latent traits in a multidimensional test.* Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Zhang, O. (2012). Observed Score and True Score Equating for Multidimensional Item Response Theory under Nonequivalent Group Anchor Test Design. *Unpublished Dissertation,* University of Florida.

APPENDIX A
TABLES

Table A-1. Ability distributions for examinee groups

Group	μ	Σ
Base Group	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Group 1	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Group 2	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} .8 & 0 \\ 0 & .8 \end{bmatrix}$
Group 3	$\begin{bmatrix} .5 \\ .5 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Group 4	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$

Table A-2. Test structure of base form unique test section (approximate simple structure)

Item	a_1	a_2	d	*MDISC*	*MDIFF*	α_1	α_2
1	0.200	0.014	0.400	0.200	-2.000	4	86
2	0.390	0.090	0.600	0.400	-1.500	13	77
3	0.589	0.114	-0.600	0.600	1.000	11	79
4	0.796	0.084	0.400	0.800	-0.500	6	84
5	0.993	0.122	0.000	1.000	0.000	7	83
6	1.169	0.270	0.000	1.200	0.000	13	77
7	1.374	0.267	1.400	1.400	-1.000	11	79
8	1.599	0.056	-0.800	1.600	0.500	2	88
9	1.799	0.063	-2.700	1.800	1.500	2	88
10	1.992	0.174	-4.000	2.000	2.000	5	85
11	0.048	0.194	0.400	0.200	-2.000	76	14
12	0.000	0.400	0.600	0.400	-1.500	90	0
13	0.000	0.600	-0.600	0.600	1.000	90	0
14	0.180	0.779	0.400	0.800	-0.500	77	13
15	0.174	0.985	0.000	1.000	0.000	80	10
16	0.249	1.174	0.000	1.200	0.000	78	12
17	0.122	1.395	1.400	1.400	-1.000	85	5
18	0.195	1.588	-0.800	1.600	0.500	83	7
19	0.126	1.796	-2.700	1.800	1.500	86	4
20	0.209	1.989	-4.000	2.000	2.000	84	6

Table A-3. Test structure of base form unique test section (complex structure)

Item	a_1	a_2	d	*MDISC*	*MDIFF*	α_1	α_2
1	0.194	0.048	0.400	0.200	-2.000	14	76
2	0.390	0.090	0.600	0.400	-1.500	13	77
3	0.593	0.094	-0.600	0.600	1.000	9	81
4	0.790	0.125	0.400	0.800	-0.500	9	81
5	0.974	0.225	0.000	1.000	0.000	13	77
6	0.946	0.739	0.000	1.200	0.000	38	52
7	1.147	0.803	1.400	1.400	-1.000	35	55
8	1.226	1.028	-0.800	1.600	0.500	40	50
9	1.474	1.032	-2.700	1.800	1.500	35	55
10	1.638	1.147	-4.000	2.000	2.000	35	55
11	0.100	0.173	0.400	0.200	-2.000	60	30
12	0.235	0.324	0.600	0.400	-1.500	54	36
13	0.401	0.446	-0.600	0.600	1.000	48	42
14	0.503	0.622	0.400	0.800	-0.500	51	39
15	0.500	0.866	0.000	1.000	0.000	60	30
16	0.084	1.197	0.000	1.200	0.000	86	4
17	0.073	1.398	1.400	1.400	-1.000	87	3
18	0.139	1.594	-0.800	1.600	0.500	85	5
19	0.374	1.761	-2.700	1.800	1.500	78	12
20	0.382	1.963	-4.000	2.000	2.000	79	11

Table A-4. Test structure of equated form unique item section (approximate simple structure)

Item	a_1	a_2	d	*MDISC*	*MDIFF*	α_1	α_2
1	0.194	0.048	0.400	0.200	-2.000	14	76
2	0.398	0.035	0.600	0.400	-1.500	5	85
3	0.599	0.031	-0.600	0.600	1.000	3	87
4	0.779	0.180	0.400	0.800	-0.500	13	77
5	0.996	0.087	0.000	1.000	0.000	5	85
6	1.193	0.125	0.000	1.200	0.000	6	84
7	1.392	0.146	1.400	1.400	-1.000	6	84
8	1.594	0.139	-0.800	1.600	0.500	5	85
9	1.796	0.126	-2.700	1.800	1.500	4	86
10	1.975	0.313	-4.000	2.000	2.000	9	81
11	0.017	0.199	0.400	0.200	-2.000	85	5
12	0.069	0.394	0.600	0.400	-1.500	80	10
13	0.063	0.597	-0.600	0.600	1.000	84	6
14	0.042	0.799	0.400	0.800	-0.500	87	3
15	0.139	0.990	0.000	1.000	0.000	82	8
16	0.000	1.200	0.000	1.200	0.000	90	0
17	0.049	1.399	1.400	1.400	-1.000	88	2
18	0.387	1.552	-0.800	1.600	0.500	76	14
19	0.188	1.790	-2.700	1.800	1.500	84	6
20	0.347	1.970	-4.000	2.000	2.000	80	10

Table A-5. Test structure of equated form unique item section (complex structure)

Item	a_1	a_2	d	*MDISC*	*MDIFF*	α_1	α_2
1	0.196	0.042	0.400	0.200	-2.000	12	78
2	0.386	0.104	0.600	0.400	-1.500	15	75
3	0.597	0.063	-0.600	0.600	1.000	6	84
4	0.796	0.084	0.400	0.800	-0.500	6	84
5	0.993	0.122	0.000	1.000	0.000	7	83
6	0.983	0.688	0.000	1.200	0.000	35	55
7	1.118	0.843	1.400	1.400	-1.000	37	53
8	1.278	0.963	-0.800	1.600	0.500	37	53
9	1.510	0.980	-2.700	1.800	1.500	33	57
10	1.509	1.312	-4.000	2.000	2.000	41	49
11	0.134	0.149	0.400	0.200	-2.000	48	42
12	0.268	0.297	0.600	0.400	-1.500	48	42
13	0.300	0.520	-0.600	0.600	1.000	60	30
14	0.424	0.678	0.400	0.800	-0.500	58	32
15	0.643	0.766	0.000	1.000	0.000	50	40
16	0.146	1.191	0.000	1.200	0.000	83	7
17	0.219	1.383	1.400	1.400	-1.000	81	9
18	0.195	1.588	-0.800	1.600	0.500	83	7
19	0.435	1.747	-2.700	1.800	1.500	76	14
20	0.313	1.975	-4.000	2.000	2.000	81	9

Table A-6. Test structure of anchor item section (approximate simple structure)

Item	a_1	a_2	d	*MDISC*	*MDIFF*	α_1	α_2
1	0.198	0.028	0.400	0.200	-2.000	8	82
2	0.397	0.049	0.600	0.400	-1.500	7	83
3	0.600	0.010	-0.600	0.600	1.000	1	89
4	0.785	0.153	0.400	0.800	-0.500	11	79
5	0.996	0.087	0.000	1.000	0.000	5	85
6	1.169	0.270	0.000	1.200	0.000	13	77
7	1.369	0.291	1.400	1.400	-1.000	12	78
8	1.576	0.278	-0.800	1.600	0.500	10	80
9	1.747	0.435	-2.700	1.800	1.500	14	76
10	1.941	0.484	-4.000	2.000	2.000	14	76
11	0.024	0.199	0.400	0.200	-2.000	83	7
12	0.056	0.396	0.600	0.400	-1.500	82	8
13	0.135	0.585	-0.600	0.600	1.000	77	13
14	0.097	0.794	0.400	0.800	-0.500	83	7
15	0.225	0.974	0.000	1.000	0.000	77	13
16	0.249	1.174	0.000	1.200	0.000	78	12
17	0.315	1.364	1.400	1.400	-1.000	77	13
18	0.195	1.588	-0.800	1.600	0.500	83	7
19	0.343	1.767	-2.700	1.800	1.500	79	11
20	0.209	1.989	-4.000	2.000	2.000	84	6

Table A-7. Repeated measure analysis results for weighted Bias and ARMSD

Statistic	Factors	Source	Partial ω^2
$Bias_w$	Between	test_str	0.02067
	Between	group	0.92557
	Between	test_str*group	0.00458
	Within	link	0.84970
	Within	link*test_str	0.00641
	Within	link*group	0.88045
	Within	link*test_str*group	0.06019
	Within	equat	0.47878
	Within	equat*test_str	0.01469
	Within	equat*group	0.46236
	Within	equat*test_str*group	0.00459
	Within	link*equat	0.00185
	Within	link*equat*test_str	0.00342
	Within	link*equat*group	0.00873
	Within	link*equat*test_str*group	0.00429
$ARMSD_w$	Between	test_str	0.00670
	Between	group	0.91944
	Between	test_str*group	0.02128
	Within	link	0.94089
	Within	link*test_str	0.03362
	Within	link*group	0.94122
	Within	link*test_str*group	0.15599
	Within	equat	0.57653
	Within	equat*test_str	0.01727
	Within	equat*group	0.58711
	Within	equat*test_str*group	0.02497
	Within	link*equat	0.38335
	Within	link*equat*test_str	0.03872
	Within	link*equat*group	0.40483
	Within	link*equat*test_str*group	0.04714

Note: link- MIRT linking methods
equate - MIRT equating methods
Group - group distribution shape
test_str - test structure

Table A-8. Weighted mean Bias for linking methods × group

Group Distribution				Linking Methods					
	Mean	SD	Cor	Min	OD	TCF	ICF	NOP	Mean
Group 1	0.0	1.0	0.0	0.32926	0.24153	0.09908	0.08588	0.58308	0.26777
Group 2	0.0	0.8	0.0	0.29248	0.25659	0.10186	0.07712	0.42949	0.23151
Group 3	0.5	1.0	0.0	-5.1191	-2.3007	-1.8481	-1.6189	-9.0351	-3.9844
Group 4	0.0	1.0	0.5	0.29875	0.64432	-0.015	-0.0637	0.74563	0.32199
Mean				-1.0497	-0.2896	-0.4155	-0.3799	-1.8192	-0.7908

Note: SD- Standard Deviation, Cor-Correlation, Min- Min's method, OD- Oshima, Davey, and Lee's Direct method, TCF- Oshima, Davey, and Lee's Test Characteristic Function method, ICF- Oshima, Davey, and Lee's Item Characteristic Function method, NOP- Reckase & Martineau's Method

Table A-9. Weighted mean ARMSD for linking methods × group

Group Distribution				Linking Methods					
	Mean	SD	Cor	Min	OD	TCF	ICF	NOP	Mean
Group 1	0.0	1.0	0.0	0.23089	0.28789	0.11543	0.11071	0.69485	0.28795
Group 2	0.0	0.8	0.0	0.20884	0.48575	0.26935	0.26063	0.55671	0.35626
Group 3	0.5	1.0	0.0	28.6865	7.52435	5.64556	4.9614	87.9546	26.9545
Group 4	0.0	1.0	0.5	0.54088	1.33343	2.00157	0.19166	1.24462	1.06243
Mean				7.4168	2.4079	2.0080	1.3811	22.6127	7.1653

Note: SD- Standard Deviation, Cor-Correlation, Min- Min's method, OD- Oshima, Davey, and Lee's Direct method, TCF- Oshima, Davey, and Lee's Test Characteristic Function method, ICF- Oshima, Davey, and Lee's Item Characteristic Function method, NOP- Reckase & Martineau's Method

Table A-10. Weighted mean Bias for equating methods × group

Group Distribution				Equating Methods			
	Mean	SD	Cor	MOSE	AOSE	ATSE	Mean
Group 1	0.0	1.0	0.0	0.23211	0.33764	0.23354	0.36066
Group 2	0.0	0.8	0.0	0.24126	0.25349	0.19976	0.2989
Group 3	0.5	1.0	0.0	-5.5469	-3.3541	-3.0522	-2.2306
Group 4	0.0	1.0	0.5	0.25488	0.40193	0.30917	0.4932
Mean				-1.2047	-0.5903	-0.5774	-0.2695

Note: SD- Standard Deviation, Cor-Correlation, MOSE- Full Information MIRT Observed Score Equating, AOSE- Unidimensional Approximation of MIRT Observed Score Equating, ATSE- Unidimensional Approximation of MIRT True Score Equating

Table A-11. Weighted mean ARMSD for equating methods × group

Group Distribution				Equating Methods			
	Mean	SD	Cor	MOSE	AOSE	ATSE	Mean
Group 1	0.0	1.0	0.0	0.23195	0.39008	0.24183	0.37277
Group 2	0.0	0.8	0.0	0.34226	0.45045	0.27605	0.37375
Group 3	0.5	1.0	0.0	40.9469	21.1754	18.7411	16.3327
Group 4	0.0	1.0	0.5	1.0374	1.34202	0.80787	0.93746
Mean				10.6396	5.8395	5.0167	4.5042

Note: SD- Standard Deviation, Cor-Correlation, MOSE- Full Information MIRT Observed Score Equating, AOSE- Unidimensional Approximation of MIRT Observed Score Equating, ATSE- Unidimensional Approximation of MIRT True Score Equating