ACT Research Report Series 2003-6

# A Comparison of Two Linking Methods for Multidimensional IRT Scale Transformations

Kyung-Seok Min

Jong-Pil Kim

ACT

December 2003

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

# A Comparison of Two Linking Methods for Multidimensional IRT Scale Transformations

Kyung-Seok Min
Jong-Pil Kim

# Abstract

Unidimensionality in the traditional IRT model has been regarded as a strong assumption. Many researchers agree that psychological/educational tests are sensitive to multiple traits, implying the need for multidimensional item response theory (MIRT). One fact that limits the application of MIRT in practice is difficulty in establishing equivalent scales based on multiple traits. Several solutions for this problem have been proposed. In this study, two MIRT linking methods, recently developed by Oshima, Davey and Lee (2000) and Li and Lissitz (2000), respectively, are investigated based on the accuracy and stability of multidimensional scale transformations under several testing conditions. Real testing outcomes, as well as simulated data, are analyzed for the comparison. The results show that Oshima et al.'s method performs well in transforming overall true test scores, and that Li and Lissitz' method has an advantage of maintaining test dimensional structures through orthogonal rotation. The limitations and cautions in using multidimensional scaling techniques are discussed.

## A Comparison of Two Linking Methods for Multidimensional IRT Scale Transformations

Traditionally, IRT models have been developed with 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 seem so simple, and many psychological and educational researchers agree that multidimensional abilities/traits come into play in test performance (Ackerman, 1991; Reckase, 1985, 1995; Traub, 1983). Most IRT linking methods have been based on unidimensional item response theory (UIRT) model. UIRT linking makes adjustments for different scales (i.e., origin and unit of scale) (Lord, 1980). When the goal is to establish comparable scores on tests that are affected by more than one dimension, however, the directions of dimensions also need to be adjusted to obtain equitable meaning. That is, multidimensional item response theory (MIRT) models are directionally indeterminant as well as scale indeterminant. Therefore, MIRT linking requires a composite transformation of rotation and scaling to derive comparable scores.

### *Purpose of the Study*

While several MIRT linking methods have been developed and share some common ground (Hirsch, 1989; Li & Lissitz, 2000; Oshima, Davey, & Lee, 2000; Thompson, Nering & Davey, 1997), each of them shows unique properties in terms of statistical characteristics and optimization criteria (i.e., what is to be minimized or maximized). Because it is not known whether different MIRT linking methods lead to the same/similar conclusions of metric transformation, careful consideration should be taken when applying any specific linking technique according to properties of each method and the goal of linking.

The purpose of this study is to evaluate two recent MIRT linking methods (i.e., Li &

Lissitz, 2000; Oshima et al., 2000) in terms of the accuracy and stability of scale transformations across various testing conditions (e.g., different sample sizes, structures of dimensionality and shapes of true ability distributions). Both simulation and real data analyses were conducted.

## MIRT and Linking Methods

### *MIRT Models*

Two types of models have been referred to in MIRT, i.e., compensatory and noncompensatory models. These are different with regard to relationships among the ability dimensions that define a person's item responses. In compensatory models (Lord & Novick, 1968; McDonald, 1967; Reckase, 1985; 1995), the proficiencies are additive in the logit, such that low ability on one trait can be compensated by high ability on other trait(s). In noncompensatory models (Sympson, 1978), a multiplication of the proficiencies bases the probability of getting an item right such that one lowest trait value among dimensions sets the upper limit of the probability. Since most research on MIRT linking has used compensatory models (partly because of estimation difficulties in the noncompensatory model) and these two types of MIRT models are indistinguishable from a practical stand point (Spray, Davey, Reckase, Ackerman, & Carlson, 1990), the compensatory model is considered in this study.

A compensatory multidimensional extension of the two-parameter logistic (2PL) model with $m$ dimensions is (Reckase, 1985; 1995)

$$P(u_{ij}=1|\mathbf{a}_i,d_i,\boldsymbol{\theta}_j)=\frac{\exp(\mathbf{a}_i'\boldsymbol{\theta}_j+d_i)}{1+\exp(\mathbf{a}_i'\boldsymbol{\theta}_j+d_i)}, \qquad (1)$$

where $P(u_{ij}=1|\mathbf{a}_i,d_i,\boldsymbol{\theta}_j)$ is the probability of a correct response for examinee $j$ on test item $i$,

$u_{ij}$ 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 $i$, $d_i$ is a parameter related to item difficulty of item $i$, and $\boldsymbol{\theta}_j$ is a vector of the $j$th examinee's abilities.

Compared with unidimensional IRT models (UIRT), multidimensional item discrimination and person ability parameters are denoted in the form of vectors rather than scalars, and the difficulty-related parameter is a composite of item difficulty and discrimination on each dimension. Interpreting MIRT discrimination parameters is analogous to UIRT parameters, but each element of the vector implies a direction in the dimensional space. The meaning of MIRT difficulty parameter is not directly equivalent to that of the unidimensional difficulty parameter because of a different parameterization. In fact, two MIRT statistics were developed to capture item characteristics corresponding to UIRT item discrimination and difficulty.

The discrimination power of a multidimensional item in the dimensional space can be defined as a function of item discrimination parameters (Ackerman, 1994; 1996, Reckase, 1985; 1995; Reckase & McKinley, 1991)

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

where $MDISC_i$ denotes the $i$th item's discrimination , $m$ is the number of dimensions in the ability space, and $a_{ik}$ is the $i$th item's discrimination on the $k$th dimension .

The multidimensional item difficulty equivalent to unidimensional difficulty is

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

where $MDIFF_i$ is the distance between the origin and the steepest point of the item response surface.

The direction of multidimensional discrimination and difficulty in the dimensional space is given by

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

where $\alpha_{ik}$ is an angle between the *k*th dimension and item vector.

As is shown in Equation (1), the probability of the correct 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 holds for a given response pattern if item parameters are transformed accordingly. In other words, the probability that an examinee gets an item right is identical when the IRT scale is changed properly. This is referred to as scale indeterminacy (Baker, 1992; Kolen & Brennan, 1995). While scale indeterminacy (location of the origin and the unit of scale) is considered in finding a proper transformation in URT linking, the rotation for the comparable reference system as well as the scale alteration has to be considered in MIRT due to multiple dimensions.

### *Linking Methods*

Even though modeling more than one dimension often improves the model fit, the use of MIRT models are limited in testing practice (Gosz & Walker, 2002; Reckase, 1997). One reason is the difficulty in finding comparable multidimensional scales across different test forms or examinee groups (Oshima et al., 2000). Several multidimensional linking methods have been proposed (i.e., Hirsch, 1989; Li & Lissitz, 2000; Oshima et al., 2000; Thompson et al., 1997); two recent MIRT linking methods, Oshima, Davey and Lee's method and Li and Lissitz's method are used in the study.

*Oshima, Davey and Lee's method.* Oshima et al.'s linking method (2000), (ODL method), is based on the common item design: a set of common items on multiple test forms are used to find a common scale. Transformations of the compensatory multidimensional model with the exponent $\mathbf{a}_i'\boldsymbol{\theta}_j + d_i$, are conducted using the following set of equations

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

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

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

where **A** is a rotation matrix and **β** is a scaling vector, and the asterisk (*) indicates transformed parameters. Two linking components, **A** and **β** are obtained by minimizing differences between test characteristic functions of common items on two test forms. Here the rotation matrix **A** does two functions: (a) determines a proper dimensional orientation (covariance/correlation), and (b) adjusts the variances of the ability dimensions. The translation vector **β** locates the origin by altering means.

The equality of the transformed exponent and the original exponent is established as

$$\mathbf{a}_i^{'*}\boldsymbol{\theta}_j^* + d_i^* = (\mathbf{a}_i'\mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\theta}_j + \boldsymbol{\beta}) + (d_i - \mathbf{a}_i'\mathbf{A}^{-1}\boldsymbol{\beta}) = \mathbf{a}_i'\boldsymbol{\theta}_j + d_i. \qquad (8)$$

Oshima et al. provided several statistical procedures to estimate transformation parameters and to evaluate linking results. They reported that the test characteristic function (TCF) method was best at finding the rotation matrix, and was also relatively good at finding the translation vector.

*Li and Lissitz's method.* Li and Lissitz's method (2000), (LL method), uses the following set of equations to transform exponential components of $\mathbf{a}_i'\boldsymbol{\theta}_j + d_i$

$$\mathbf{a}_i^* = k\mathbf{a}_i'\mathbf{T}, \qquad (9)$$

$$d_i^* = d_i + \mathbf{a}_i'\mathbf{Tm}\text{ , and} \qquad (10)$$

$$\boldsymbol{\theta}_j^* = (1/k)(\mathbf{T}^{-1}\boldsymbol{\theta}_j - \mathbf{m}), \qquad (11)$$

where **T** is an orthogonal rotation matrix for direction, **m** is a translation vector for location, and $k$ is a central dilation constant for unit. The rotation matrix, **T** is obtained by orthogonal Procrustes solutions, and **m** and $k$ are calculated by minimizing differences between base and transformed discriminations and difficulties of common items, respectively. The equality of the transform components and the original components is established as

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

Note that Equations (5) to (7) are mathematically equivalent to Equations (9) to (11) except for pre-multiplication or post-multiplication of the rotation matrix.

Li and Lissitz tried to provide a multidimensional linking method by taking into account three linking components, i.e., rotation, translation, and central dilation (refer to Schönemann, 1966; Schönemann & Carrol, 1970). It is straightforward in that three linking components can provide useful information to compare different test forms or different examinee groups even though initial procedures were developed from the anchor item design. While the ODL method deals with dimensional direction and unit change at once in the rotation matrix, the LL method splits these two components into the rotation matrix and the central dilation. Here, 'central' means that unit changes are assumed to be similar across dimensions such that one scalar ($k$) can cover overall unit changes.

## Method

### *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 (Bolt, 1999; Davey, Nering, & Thompson, 1997). Since we can know true parameters in the simulation study, it is easier to compare true parameters with their estimates.

*Linking design and specification of the item response model.* Two test forms sharing a set of common items were used, the so-called common item design. Suppose one form is the base test and the other is the linked test, and each of them include common items and unique items. The linked test scores need to be converted into base test scores. The common item set was used as a way to find a comparable test scale. In order to calibrate item and ability estimates, a compensatory two-dimensional 2PL model was used as Equation (1).

*Generation of true item parameters and item response patterns.* Item parameters were drawn from probability distributions of where the ranges were determined by the specification of dimensional structures. Two types of item dimensional structures were investigated: approximate simple structure (APSS) and mixed structure (MS). These two structures have been discussed as being more realistic than the simple structure (SS) that is an ideal one (Kim, 1994; Roussos, Stout, & Marden, 1998). APSS means that each item highly but not fully loads on one of the dimensions. In other words, a set of items has high discriminations on the same dimension. However, in reality test items likely measure some composite of dimensions as well as pure dimensions. MS refers to a test that measures both relatively pure trait dimensions and composites of dimensions. For the present simulation, APSS was constructed by two sets of items. One set of items loaded mainly on the first dimension and the other set loaded on the second dimension. In MS, there were four sets of items. Two sets loaded heavily on one of the two dimensions and the remaining two sets were loaded to composites of the two dimensions. These two-dimensional structures for the 20 common items are illustrated in Figure 1.

In order to define item parameters, fixed values of MDISCs and MDIFFs generated by Roussos et al. (1998) are given in Table 1.

**TABLE 1**

**Five MIRT Discrimination and Difficulty Levels**

| Level | MDISC | MDIFF |
|---|---|---|
| 1 | 0.4 | −1.5 |
| 2 | 0.8 | 1.0 |
| 3 | 1.2 | 0.0 |
| 4 | 1.6 | −1.0 |
| 5 | 2.0 | 1.5 |
| Mean | 1.2 | 0.0 |

These two sets of MIRT characteristics were selected because they are realistic, cover item features usually found on a test, and they do not relate dimensionality and item difficulty levels (Roussos et al., 1998). There are five sets of MDISCs and MDIFFs, so four of the 20 common items had each set. Discrimination and difficulty-related parameters were determined by Equations (2), (3), and (4). A set of item parameters that were used for the present simulation is given in Table 2.

**TABLE 2**

**Item Parameters of 20 Common Items**

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

Given the MIRT item parameters, the response probability $P_{ij}$ was computed for each examinee. Then $P_{ij}$ was compared to a uniform random value $P^*$ where $0 \leq P^* \leq 1$. A binary item score of $x_{ij} = 1$ was assigned when $P_{ij} > P^*$. Otherwise, a score of $x_{ij} = 0$ was assigned.

*Specification of examinee ability distributions.* Five multivariate normal distributions with various means and variances/covariances were considered for examinee true abilities (Table 3). Different distributions reflected different examinee groups across multiple test forms. The distribution of Group 1 is the default ability distribution (standardized bivariate normal distribution) assumed in MIRT calibration programs (e.g., NOHARM, Fraser, undated).

### TABLE 3

**Ability Distributions for Five Examinee Groups**

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 |
|---|---|---|---|---|---|
| Mean, Variance/ Covariance | $\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ | $\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & .5\\.5 & 1\end{bmatrix}$ | $\begin{bmatrix}.5\\.5\end{bmatrix}, \begin{bmatrix}1 & .5\\.5 & 1\end{bmatrix}$ | $\begin{bmatrix}.5\\.5\end{bmatrix}, \begin{bmatrix}.8 & .4\\.4 & .8\end{bmatrix}$ | $\begin{bmatrix}.5\\.5\end{bmatrix}, \begin{bmatrix}1.2 & .5\\.5 & .8\end{bmatrix}$ |
| Correlation | .00 | .50 | .50 | .50 | .51 |

For Group 2, true abilities were assumed to have moderate correlation ($\rho$=0.5). Because a calibration program assumes independent dimensions, the direction of dimensions becomes arbitrary and sample specific. It should be noted that item discrimination estimates reflect the correlation among ability dimensions when independent dimensions are calibrated. In Group 3, a different mean vector from the default of zero for both dimensions was considered. For the last two groups, variances of abilities varied, but the correlations of two dimensions were maintained at about .5.

*The number of examinees.* Generally, 2000 or more examinees seem to be sufficient for MIRT calibration. In order to evaluate the stability of linking for small sample sizes, relatively small numbers of examinees (500 and 1000) along with the recommended size of 2000 were used.

*Number of replications.* With two-dimensional structures, three sample sizes and five ability distributions, there were thirty combinations of simulation conditions. Fifty test response patterns were generated for each combination.

*Calibration and linking.* Before conducting linking, item parameters were estimated with NOHARM. Following the common item design, estimates of common items were transformed to the initial item parameters that were provided in Table 2. Two sets of the common item estimates through two independent calibrations were used for the base and the linked scale separately. In the simulation study, the item parameters were used as the base scale, and the estimates under various conditions were used as the linked scale.

The two linking methods, the ODL and LL linking methods, were compared based on how closely item estimates were transformed into item parameters, i.e., degree of parameter recovery. For each method, there were several sub-procedures, which resulted in slightly different transformations. One relatively best sub-procedure was selected for each method: the test characteristic function procedure (TCF) for the ODL method and the composite procedure of orthogonal Procrustes solutions for the LL method.

Several computer programs were used in the simulation study. In order to generate ability distributions that are multivariate normal with given means and variances/covariances, GENDAT5 (Thompson, undated), a modified version of NOHARM (Thompson, 1996; Fraser, undated) was used. IPLINK (Lee & Oshima, 1996) and MDEQUATE (Li, 1996) were run to implement the two linking methods, respectively.

*Evaluation criteria.* Although the ODL and LL linking methods easily apply to other linking designs, they were originally developed for the common item design. In the IRT framework, one of the evaluation criteria for the common item linking is how small the differences are between base estimates and transformed estimates. 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 (bias), and (b) how much differences fluctuate (root mean square error, RMSE) across common items. Bias and RMSE were computed by

$$\sum_{i=1}^{I} \frac{(\hat{a}_{1i}^{*} - a_{1i})}{I}, \text{ and} \tag{13}$$

$$\left( \frac{\sum_{i=1}^{I} (\hat{a}_{1i}^{*} - a_{1i})^{2}}{I-1} \right)^{1/2}, \tag{14}$$

respectively, where $a_{1i}$ is the discrimination parameter on the first dimension of item $i$ , $\hat{a}_{1i}^{*}$ is the transformed discrimination, and *I* is the number of common items.

The same formulas were applied to other item characteristics; discrimination on the second dimension ($a_2$) and difficulty related parameters (d). As each item has three parameters and transformed values, there are three sets of bias and RMSE for each replication.

Because two linking methods were applied to the same test response patterns, the repeated measures analysis of variance was used to detect effects of simulation conditions and linking methods on bias and *ln (RMSE)*.[1] The model is

$$Bias(a_1)_{lings} = \mu + \beta_n + \gamma_g + \lambda_s + \gamma\lambda_{gs} + \pi_{i(ngs)} + \alpha_l + \alpha\beta_{ln} + \alpha\gamma_{lg} + \alpha\lambda_{ls} + \alpha\gamma\lambda_{lgs} + e_{li(ngs)}, \tag{15}$$

where $Bias(a_1)_{lings}$ is 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; $\mu$ is overall mean in population; $\beta_n$ is

[1] In order to obtain a more desirable distribution (normality), a natural logarithm was taken for RMSEs.

effect of *n*th sample size (500, 1000 and 2000); $\gamma_g$ is effect of *g*th group (Groups 1 to 5); $\lambda_s$ is effect of *s*th dimensional structure (APSS and MS); $\gamma\lambda_{gs}$ is interaction effect of group and structure; $\pi_{i(ngs)}$ is effect of *i*th iteration within *n*th sample size, *g*th group and *s*th structure (iteration 1 to 50); $\alpha_l$ is effect of *l*th linking method (the ODL and LL methods); $\alpha\beta_{\mathrm{ln}}$ is interaction effect of linking method and sample size; $\alpha\gamma_{\mathrm{lg}}$ is interaction effect of linking method and group; $\alpha\lambda_{ls}$ is interaction effect of linking method and dimensional structure; $\alpha\gamma\lambda_{\mathrm{lg}s}$ is interaction effect of linking method, group, and dimensional structure; and $e_{li(ngs)}$ is interaction effect of linking method and iteration within *n*th size, *g*th group and *s*th structure.[2]

In the model of Equation (15), there are three between-factors; sample size, group and structure. The interaction term of between-factors was selected based on the initial examination of full model results. Also there is one within-factor, linking method, and others are interaction terms of between- by within-factors. Equation (15) is the model for the bias of the first dimensional discrimination and the same model applies to other bias measures and log transformed RMSE for all three item parameters. After conducting statistical tests, patterns of biases and RMSEs were examined in detail across simulation conditions and linking methods.

*Real data analysis.* Simulation data have advantages in that they can clarify which factor/condition(s) leads to favorable or unfavorable results, because one knows the true model and parameters. However, statistical models including measurement models emulate real situations at best, they are not reality itself. So the overall evaluation of simulation studies

[2] Statistical tests of the repeated measures analysis of variance model are based on the symmetry condition:1) the variance-covariance matrix of transformed variables used to test effects has covariances of zero and equal variances (sphericity), and 2) the variance-covariance matrix must be equal for all levels of between subject factors (homogeneity).

depend on how plausible the assumed conditions and following resultant data are. One way to scrutinize a simulation study is to compare its results with real data, and see if both lead to consistent conclusions.

For this purpose, actual test response data was analyzed. Test items for the real data were from a readiness test, designed to measure two distinguishable traits. There were different readiness test forms, and all forms consisted of 40 items; 25 for reading and 15 for mathematics. Among 40 items, 12 reading items and 10 mathematics items were common items across different test forms. Sample sizes for each form were relatively small. In real data analyses, the differences between base test item estimates and the transformed estimates were evaluated. In addition to the item level comparison, the differences of true scores using the test response surfaces, were examined for two linking methods.

## Results

### *Simulation Study*

Linking errors for each replication of 20 common items were summarized by two statistics, mean and standard deviation of differences between transformed values and item parameters. These two statistics were considered as indicators of the quality of linking for each replication. After finding significance of multivariate statistics in Equation (15), univariate tests for six dependent variables were implemented. The results are provided in Table 4. The statistical test results indicated that the effects of linking method depend on simulation conditions (i.e., significant results for within-factor interaction terms). In addition to the interaction of between- by within-factors, three main factors of the simulation conditions had significant effects on linking bias and log transformed RMSE. The bias of the first dimensional discrimination was most sensitive to simulation conditions and linking methods while bias of difficulty was least sensitive. In general, the results of the repeated measures ANOVA showed that the soundness of the linking results depended on test conditions and linking methods.

**TABLE 4**

**Repeated Measures ANOVA (F values)**

| Source of Variation | Bias ($a_1$) | Bias($a_2$) | Bias(d) |
|---|---|---|---|
| Between Factor | | | |
| Sample Size | 152.41** | 148.21** | 33.72** |
| Group | 82.91** | 87.83** | 2.16 |
| Dimensional | 274.40** | 229.05** | 15.35** |
| Group*Structure | 16.11** | 15.67** | 2.76 |
| Within Factor | | | |
| Linking Method | 653.21** | 69.06** | 329.44** |
| Link*Size | 117.53** | 127.42** | 17.75** |
| Link*Group | 25.41** | 29.50** | 1.31 |
| Llink*Structure | 89.56** | 84.65** | 7.24** |
| Link*Group*Structu | 7.32** | 6.54** | .71 |
| | LN RMSE ($a_1$) | LN RMSE ($a_2$) | LN RMSE(d) |
| Between Factor | | | |
| Sample Size | 442.84** | 459.43** | 327.62** |
| Group | 153.42** | 180.64** | 8.69** |
| Dimensional | 2.88 | .47 | .20** |
| Group*Structure | 1.21 | 1.04 | 3.30* |
| Within Factor | | | |
| Linking Method | 226.53** | 238.59** | 1811.99** |
| Link*Size | 128.05** | 156.70** | 32.03** |
| Link*Group | 75.23** | 71.99** | 1.00 |
| Llink*Structure | 25.37** | 35.59** | 5.98** |
| Link*Group*Structu | 4.12** | 4.52** | 5.07** |

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

To evaluate the behavior of the two MIRT linking methods across simulation conditions, two summary statistics were plotted in Figures 2 through 7. Each data point of lines represents the average of linking errors of 50 replications in terms of bias and RMSE. Note that, for example, APSS1 indicates the Group 1 with APSS items. In general, one can notice that the ODL method was less biased and more stable than the LL method for two discrimination parameters and the LL method did better transformations for difficulty estimates than the ODL method. More detailed results follow:

1. There was inconsistency in Figures 2 and 3 that as the sample size increases, linking became more biased with the LL method. For the ODL method, however, larger samples reduced the linking bias consistently. Also, in the ODL method, biases are relatively small and stable across different sample sizes compared with the LL method.
2. Figure 4 shows that the difficulty estimates are over-transformed in the LL method, while under-transformed with the ODL method.
3. Figures 5 and 6 indicate that the ODL method showed larger RMSEs for two discrimination parameters than the LL method when the sample size was small. But as sample size increased, the ODL method generated more stable transformations than the LL method.
4. Figures 2, 3, 5 and 6 show that linking bias and RMSE of the LL method for discrimination parameters had relatively big differences between Group 1 and 2 (whether or not dimensions are correlated).
5. Figure 7 shows that the transformed difficulty estimates of the LL method were more stable than those of the ODL method.

In sum, the simulation study revealed that the accuracy (bias) and stability (RMSE) of linking depended much on the selection of linking method, sample size, examinee ability distribution, and dimensional structure. Even though there was some inconsistency in biases, especially for the LL method, as the number of examinees increased, metric transformations became less biased and more stable. Generally speaking, the ODL method did better transformations for discrimination parameters and the LL method did better transformation for difficulty related parameters.

### *Real Data*

As a real data example, two readiness test forms were analyzed. Because the test was developed to measure two distinguishable abilities (i.e., reading and mathematics), the two dimensional model was used to analyze the data. The lower asymptote parameter was ignored because traditional item difficulties of all the items were around .9 (about 90% of examinees get each item right). One form was treated as the base form and the other as the linked form. Originally there were 22 common items across test forms, but all examinees in the sample responded correctly to one of the reading items, so that item was removed. Item parameter estimates after a varimax rotation were used in order to clarify the dimensional structure. Estimates of 21 common items (11 items for reading and 10 for mathematics) are provided in Table 5.

After the linked form was transformed onto the base form scale by using the ODL method and the LL method, the two transformed values were compared with the item estimates of the base form. Differences between transformed estimates and base estimates are illustrated in Table 6. It shows that the ODL method transformed discrimination estimates of the linked form onto the base form more closely than the LL method did, and that the LL method worked better for difficulty estimates, which confirmed the simulation results.

## TABLE 5

**Item Parameter Estimates of Two Readiness Test Forms**

| Item | | Base Test (n=190) | | | Linked Test (n=199) | | |
|---|---|---|---|---|---|---|---|
| | | $a_1$ | $a_2$ | d | $a_1$ | $a_2$ | d |
| Reading | 1 | 0.91 | 0.24 | 2.66 | 0.84 | 0.52 | 11.56 |
| | 2 | 0.91 | 0.75 | 3.32 | 0.73 | 0.48 | 3.98 |
| | 3 | 2.38 | 2.44 | 7.22 | 0.76 | 0.28 | 3.20 |
| | 4 | 1.01 | -0.05 | 2.37 | 0.72 | 0.35 | 3.05 |
| | 5 | 0.61 | 0.34 | 1.68 | 0.66 | 0.29 | 2.23 |
| | 6 | 0.59 | 0.81 | 1.77 | 0.41 | 0.41 | 1.43 |
| | 7 | 0.42 | 0.67 | 1.59 | 0.63 | 0.25 | 1.58 |
| | 8 | 0.69 | 0.44 | 1.48 | 0.66 | 0.28 | 1.80 |
| | 9 | 1.73 | 0.52 | 2.64 | 0.60 | 0.22 | 1.41 |
| | 10 | 0.90 | 0.60 | 1.73 | 0.61 | 0.52 | 1.85 |
| | 11 | 0.87 | 0.36 | 1.03 | 0.49 | 0.28 | 1.04 |
| Math | 12 | 1.03 | 1.35 | 3.19 | 0.13 | 0.80 | 2.20 |
| | 13 | 1.30 | 0.93 | 2.67 | 0.20 | 0.88 | 3.42 |
| | 14 | 1.14 | 2.02 | 3.97 | 0.34 | 0.92 | 6.72 |
| | 15 | 0.67 | 1.00 | 2.40 | 0.49 | 0.52 | 2.10 |
| | 16 | 0.63 | 1.26 | 2.32 | 0.23 | 0.88 | 3.15 |
| | 17 | 0.43 | 1.21 | 1.28 | 0.27 | 0.77 | 1.76 |
| | 18 | 0.81 | 0.94 | 1.74 | 0.51 | 0.48 | 1.51 |
| | 19 | 0.86 | 0.40 | 0.99 | 0.11 | 0.58 | 0.63 |
| | 20 | 0.56 | 0.83 | 1.39 | 0.42 | 0.61 | 1.33 |
| | 21 | 0.77 | 2.13 | 1.61 | 0.22 | 0.69 | 0.71 |
| Mean | | 0.91 | 0.91 | 2.34 | 0.48 | 0.52 | 2.70 |
| SD | | 0.44 | 0.63 | 1.33 | 0.22 | 0.22 | 2.39 |

## TABLE 6

**Mean Differences Between Base Estimates and Transformed Values on the Readiness Test**

| | $a_1$ | $a_2$ | d |
|---|---|---|---|
| ODL method | –.32 | –.34 | –.38 |
| LL method | –.45 | –.48 | –.20 |

When the two linking methods' results were compared on the true score scale (test response surface), differences between the two sets of true scores (transformed score minus true base test score) were calculated for limited points (49 cells) of a two-dimensional ability space (7 by 7). These difference scores, along with true score estimates of the base form, are presented in Figure 8, which shows that the two linking methods had different patterns of linking errors. For the ODL method, there were relatively moderate gaps between base test scores and transformed scores when examinees had lower or higher scores, cells of the low left corner or center of the first panel. However, the LL linking showed relatively large and moderate differences when students had lower test scores.

## Summary and Discussion

The present study evaluated two MIRT linking methods based on the compensatory two-dimensional 2PL model and the common item design. The repeated measures ANOVA results showed that selection of linking method had a statistically significant impact on linking errors. When degrees of the recovery of parameters were quantified in terms of bias and RMSE, the LL method worked better than ODL method for difficulty parameters, and the ODL worked better for two discrimination parameters. While the sample of 500 examinees showed unreliable results, mainly due to instability of estimates, but linking results with 1000 examinees showed somewhat acceptable outcomes compared with metric transformations with the recommended sample size, 2000, in MIRT.

In real data analysis of two readiness test forms with less than 200 examinees, again the LL method slightly outperformed the ODL method for difficulty parameters and the ODL method worked better for discrimination parameters. However, the comparison of test response surfaces (e.g., true score at 49 ability points) revealed different error regions for the two linking results even though the ODL method showed fewer differences on average. While the ODL method had moderate linking errors to both lower and higher scored examinees, the LL method generated large and moderate errors to low level students.

Major statistical differences between the two linking methods could be found in linking components and optimization criteria. The ODL linking method consists of two components, the rotation matrix and the translation vector, while the LL method includes the central dilation constant in addition to the previous two components. In some sense, the rotation matrix of the ODL method might be considered as a composite of the rotation matrix and the dilation constants of the LL method because it is supposed to alter the variance/covariance of the initial distribution. More distinguishable differences between the two methods are optimization procedures to estimate linking components. The ODL method is an expansion of the UIRT linking framework (Stocking & Lord, 1983) such as minimizing the differences between two true score surfaces. The LL method adopts traditional factor analysis techniques to obtain comparable scales such as the orthogonal Procrustes solutions that are supposed to minimize differences between two matrices through a composite transformation.

Real data analysis showed that the two linking methods contained different amounts of linking error on an ability dimension. As in the simulation study, the ODL method outperformed the LL method. The two methods showed different patterns in the true score error. For example, if any critical decision were made around lower test score for an examinee that takes the linked form of the readiness test, the ODL linking method would be more conservative than the LL method. On the other hand, for moderate or higher test scores, the two linking methods perform equally well or the LL method works better in some occasions. Thus, the selection of linking method depends on the purpose of the linking in that it would be a situational specific decision.

In general, a linking procedure requires individual judgements that are made by the individuals who are doing the linking. The judgment should be informed by practical testing issues and statistical characteristics of linking techniques. MIRT and linking of multidimensional test scales are relatively new, uncertain areas compared with a huge volume of research on UIRT linking. Further research is needed on various issues to make MIRT linking more applicable, such as evaluation criteria, small sample size, the number of common items and non-normal ability distributions.

## References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*, 12-24.

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

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

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

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education, 12*, 383-407.

Davey, T., Nering, M. L., and Thompson, T. (1997). *Realistic simulation of item response data.* ACT Research Report Series.

Fraser, C. (undated). *NOHARM*: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.

Gosz, J. K., and Walker, C. M. (2002). *An empirical comparison of multidimensional item response data using TESTFACT and NOHARM.* Paper presented at the annual meeting of the National Council for Measurement in Education. New Orleans, Louisiana.

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

Kim, H. (1994). *New techniques for the dimensionality assessment of standardized test data.* Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Kolen, M. J., and Brennan, R. L. (1995). *Test Equating: Methods and Practices.* New York: Springer.

Lee, K., and Oshima, T. C. (1996). *IPLINK*: Multidimensional and unidimensional item parameter linking 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 linking. *Applied Psychological Measurement, 24,* 115 – 138.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading , MA: Addison-Wesley.

McDonald, R. P. (1967). *Nonlinear factor analysis* (Psychometric Monographs, No. 15). Iowa City: Psychometric Society.

Mislevy, R. J., and Bock, R. D. (1990). *BILOG-3: Item Analysis and Test Scoring with Binary Logistic Models* [Computer Software]. Mooresvill, IN: Science Software.

Oshima, T. C., Davey, T. C., and Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, 37(4), 357-373.

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

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. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

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.

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. (1966). A Generalized solution of the orthogonal Procrustes problem. *Psychometrika, 31,* 1-10.

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

Spray, J. A., Davey, D. C., Reckase, M. D., Ackerman, T. A., and Carlson, J. E. (1990). *Comparison of Two Logistic Multidimensional Item Response Theory Models* (ACT research report series ONR90-8). Iowa City, IA: ACT inc.

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. (Undated). *GENDAT5*: A computer program for generating multidimensional item response data.

Thompson, T. (1996). *NOHARM21*: NOHARM (C. Fraser, undated) converted to Windows.

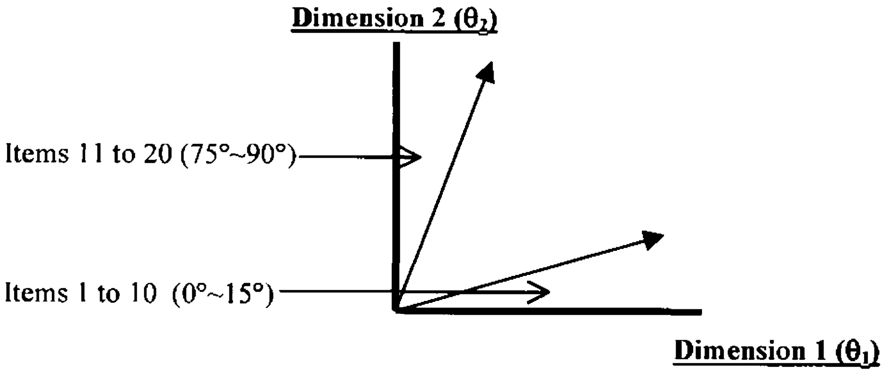
Thompson, T., Nering, M., & Davey, T. (1997, June). *Multidimensional IRT Scale Linking without Common Items or Common Examinees.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

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

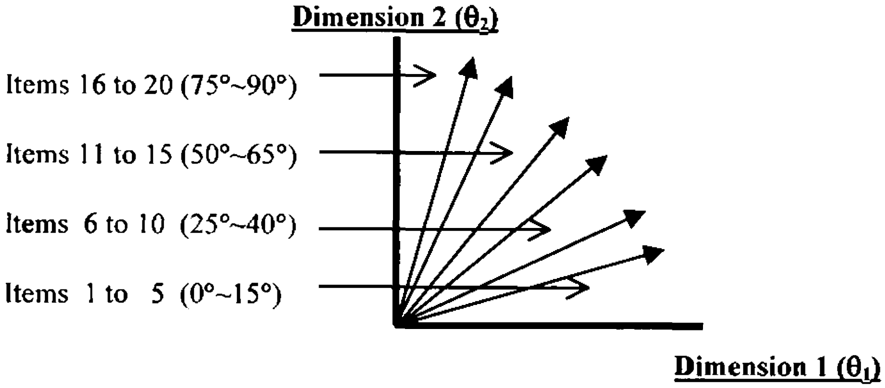
Wingersky, M. S., Barton, M. A., and Lord, F. M. (1982). *LOGIST V User's Guide.* Princeton, NJ: Educational Testing Service.

*FIGURE 1.* **Two Dimensional Structures for Simulation Data**

Note: All angles are defined from the dimension 1.



(a) Approximate Simple Structure



(b) Mixed Structure

*FIGURE 2.* **Bias for the First Dimension Discrimination ($a_1$)**

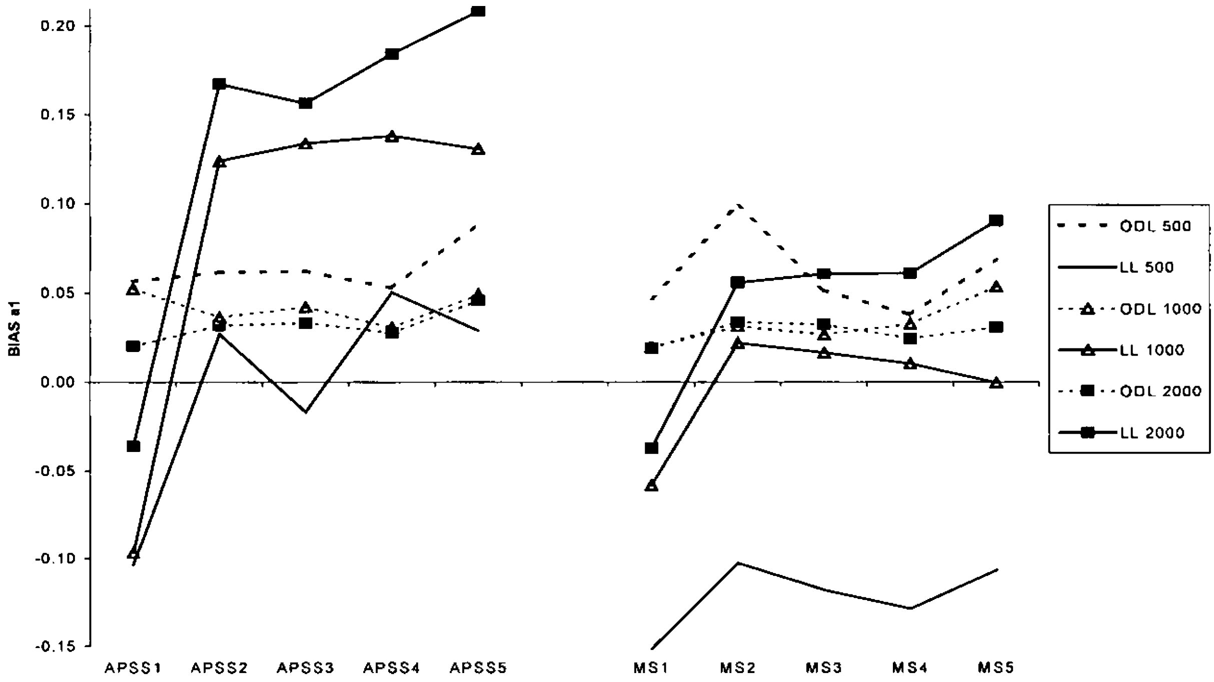


*FIGURE 3*. **Bias for the Second Dimension Discrimination ($a_2$)**

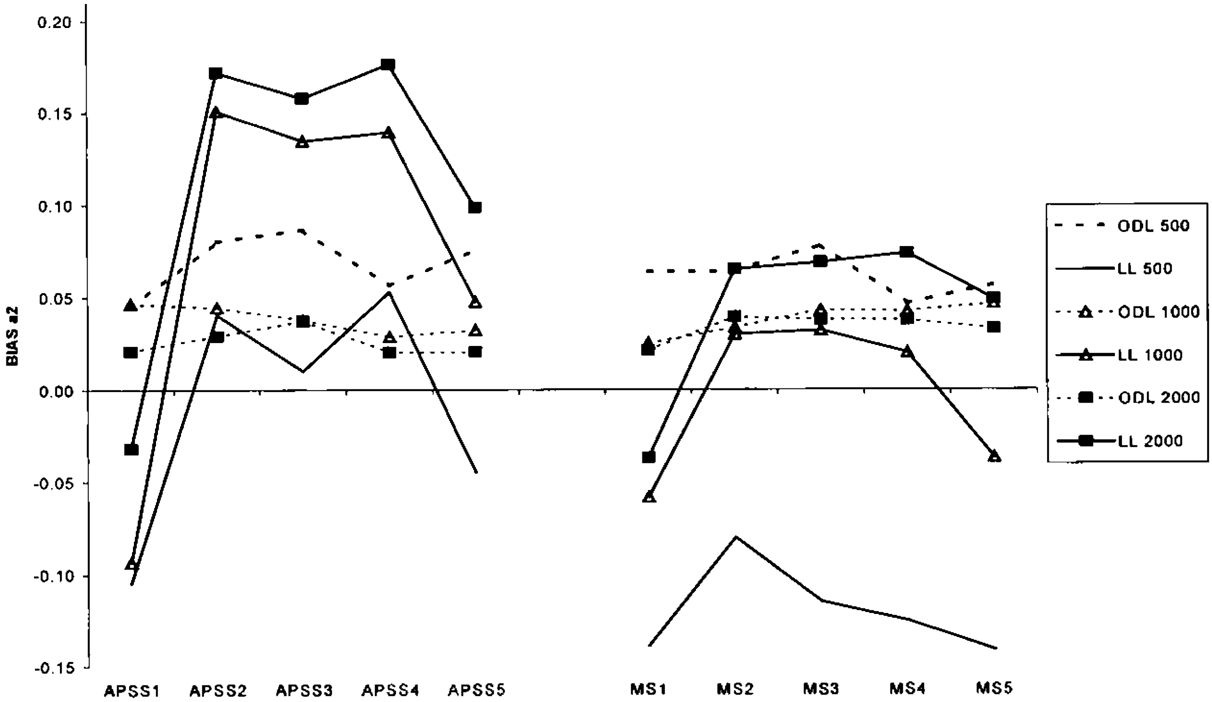


*FIGURE 4.* **Bias for the Difficulty (d)**

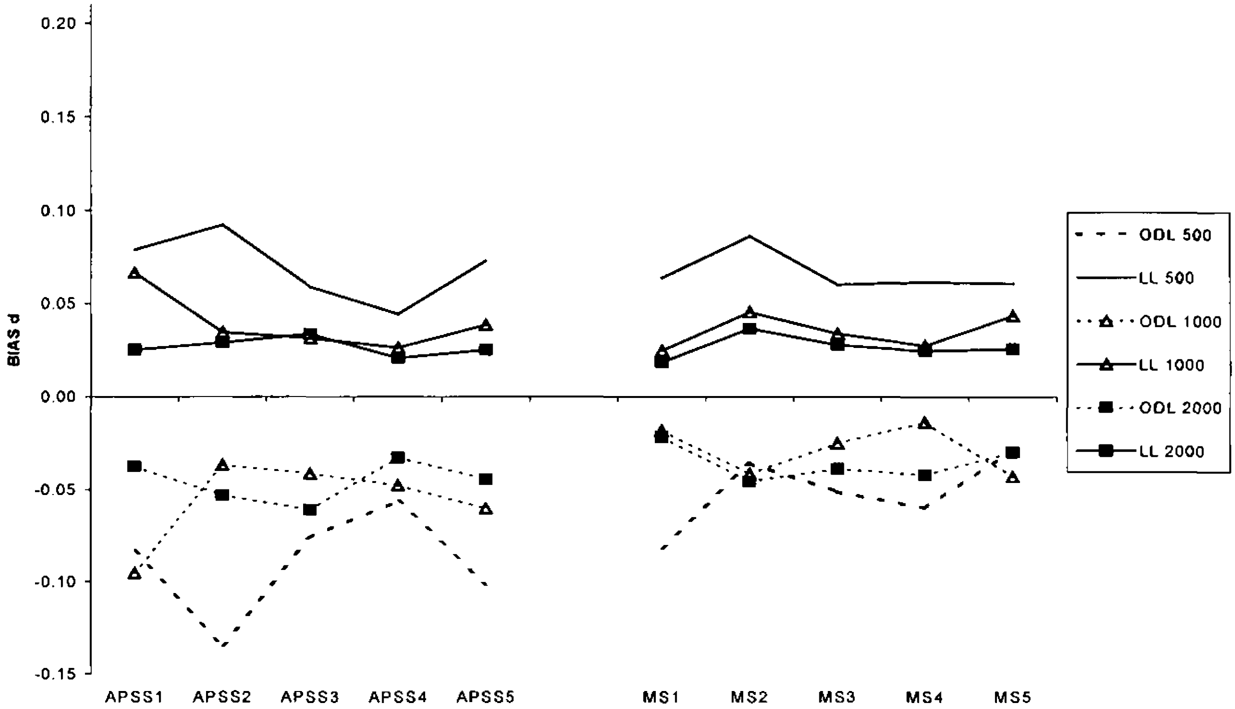


*FIGURE 5.* **RMSE for the First Dimension Discrimination ($a_1$)**

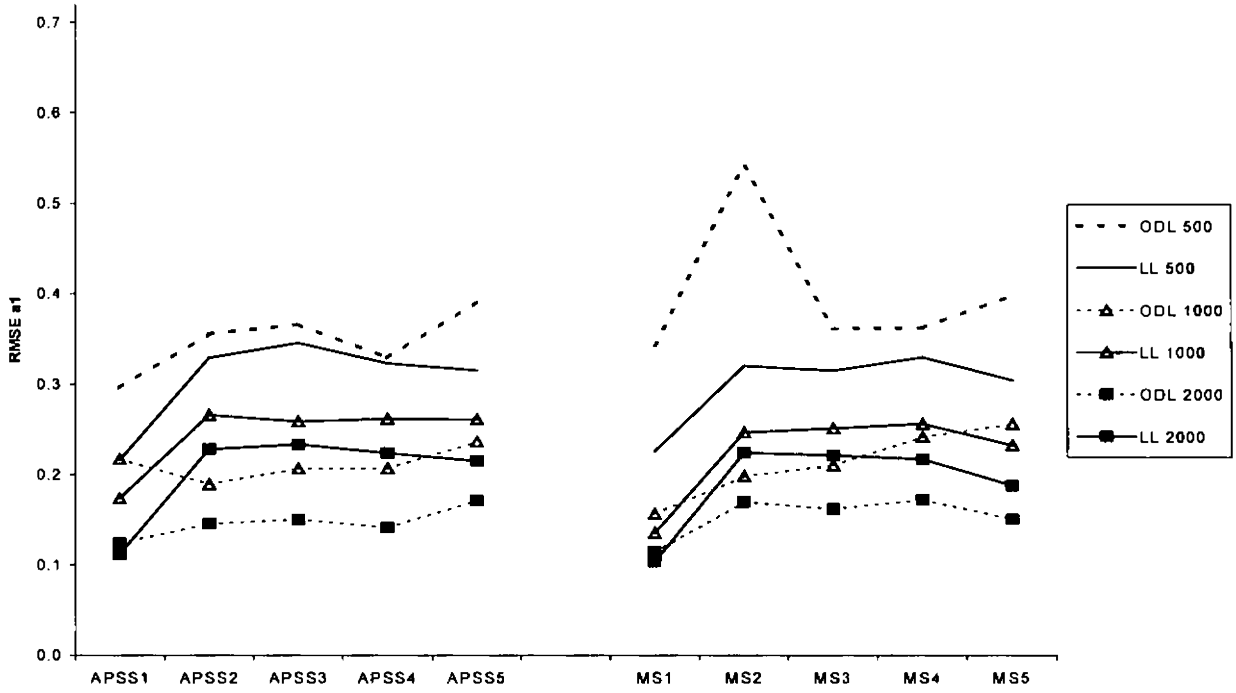


*FIGURE 6.* **RMSE for the Second Dimension Discrimination ($a_2$)**

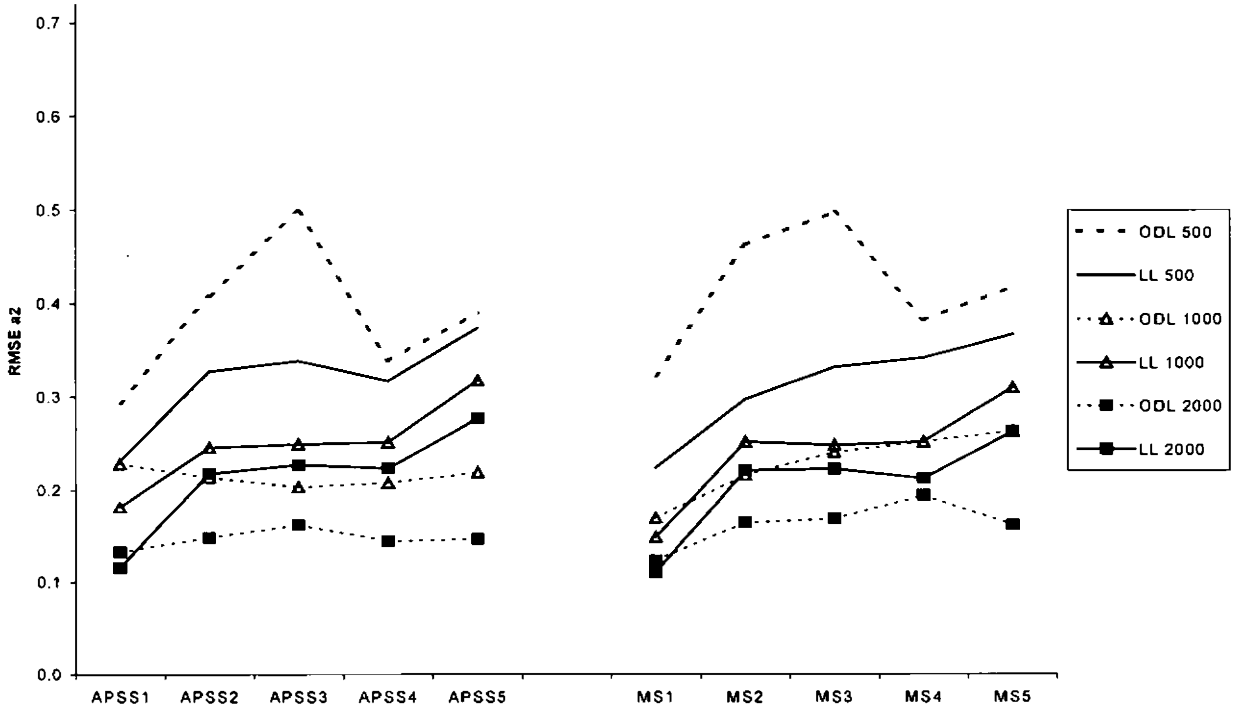


*FIGURE 7.* **RMSE for the Difficulty (d)**

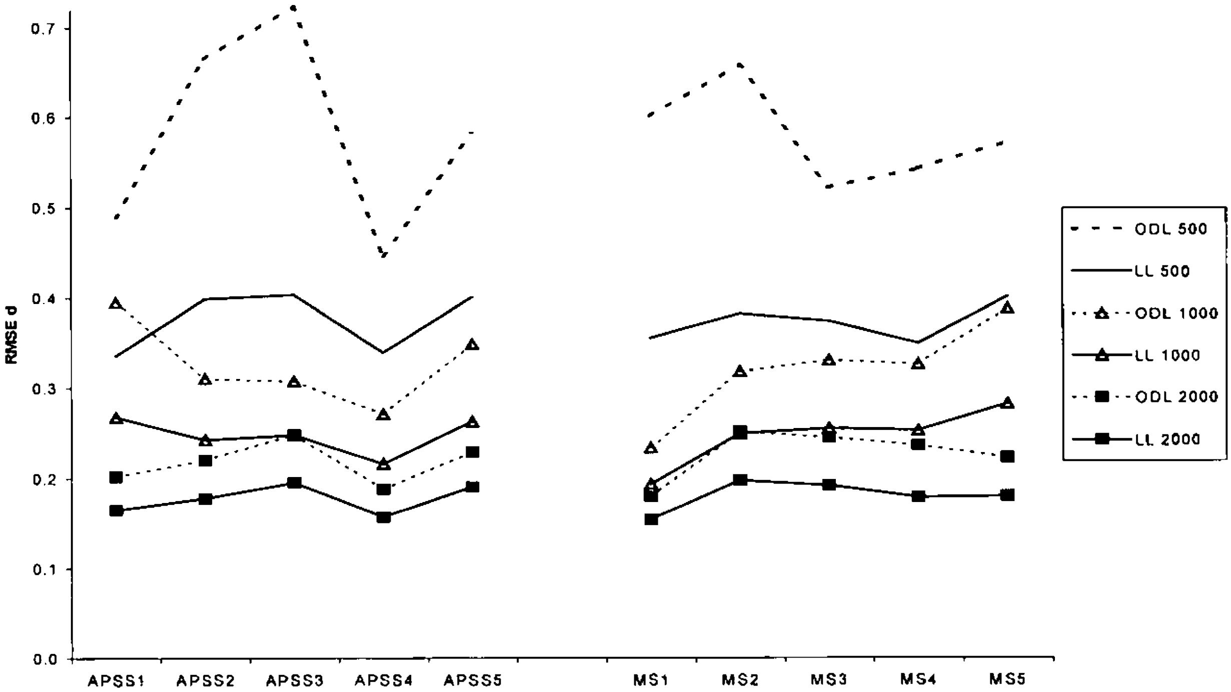


***FIGURE 8.*** **Differences of True Scores on the Base and the Transformed Scales**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 0.73 | -0.01 | -0.32 | -0.23 | -0.12 | -0.06 | -0.03 |
| 2 | -0.42 | -0.73 | -0.85 | -0.53 | -0.27 | -0.13 | -0.06 |
| 1 | -1.17 | -1.71 | -1.79 | -1.13 | -0.58 | -0.27 | -0.12 |
| $\theta_2$ 0 | 0.68 | -1.45 | -2.58 | -2.04 | -1.16 | -0.56 | -0.26 |
| -1 | 2.87 | 1.24 | -1.43 | -2.14 | -1.55 | -0.88 | -0.44 |
| -2 | 2.66 | 2.51 | 1.12 | -1.03 | -1.30 | -0.56 | -0.03 |
| -3 | 2.02 | 2.01 | 1.83 | 0.61 | -0.40 | -0.28 | 0.26 |
| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |

$\theta_1$

(a) ODL method

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 0.15 | -0.53 | -0.64 | -0.41 | -0.22 | -0.12 | -0.07 |
| 2 | 0.09 | -0.66 | -0.95 | -0.66 | -0.39 | -0.22 | -0.14 |
| 1 | 0.61 | -0.63 | -1.33 | -1.07 | -0.70 | -0.44 | -0.28 |
| $\theta_2$ 0 | 3.46 | 0.75 | -1.21 | -1.51 | -1.18 | -0.83 | -0.58 |
| -1 | 6.18 | 4.34 | 0.99 | -0.78 | -1.23 | -1.21 | -1.01 |
| -2 | 5.75 | 5.99 | 4.38 | 1.33 | -0.26 | -0.70 | -0.83 |
| -3 | 4.45 | 5.24 | 5.44 | 3.82 | 1.61 | 0.19 | -0.52 |
| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |

$\theta_1$

(b) LL method

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 16.97 | 19.38 | 20.57 | 20.90 | 20.97 | 20.99 | 21.00 |
| 2 | 15.61 | 18.54 | 20.29 | 20.81 | 20.95 | 20.99 | 21.00 |
| 1 | 13.23 | 17.04 | 19.68 | 20.62 | 20.89 | 20.97 | 20.99 |
| $\theta_2$ 0 | 8.19 | 13.66 | 18.04 | 20.01 | 20.68 | 20.89 | 20.96 |
| -1 | 3.22 | 7.77 | 13.79 | 17.70 | 19.58 | 20.43 | 20.77 |
| -2 | 1.47 | 3.73 | 8.03 | 13.52 | 16.96 | 18.64 | 19.59 |
| -3 | 0.86 | 2.26 | 4.57 | 8.69 | 13.01 | 16.02 | 17.85 |
| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |

$\theta_1$

(c) True score distribution on the base test