International Conference on Education and Educational Psychology (ICEEPSY 2012)

The Effect of Proportion Common Item’s with Mixed Format Test on Multidimensional Item Response Theory Linking

Siwatussana Suksuwan*, Putcharee Junpengb, Sungworn Ngudgratokec, Panat Guayjarernpanishkd

*Ph.D. Student in Measurement And Evaluation Khon Kaen University, Khon Kaen 40002, Thailand
bFaculty of Education Khon Kaen University, Khon Kaen 40002, Thailand
cSchool of Education Sukhothai Thammathirat Open University, Nonthaburi 11120, Thailand
dFaculty of Science Ubon Ratchathani University, Ubon Ratchathani 34190, Thailand

Abstract

Numerous assessments contain a mixture of multiple choice and constructed response item types, which are found to measure more than one trait. Thus, there is a need for multidimensional dichotomous and polytomous item response theory modeling solutions, including multidimensional linking scores. Common items are most important for linking scores. Practitioners need empirical data to inform their selection of the number common items’ for linking scores. Previous research conducted in tests composed exclusively of multiple-choice items demonstrated that common-item sets should be representative of the overall test. The extent to which the number of common items in a mixed-format test differed measured different aspects of the construct. This research explored the effects of proportion common items’ with mixed format test on the multidimensional item response theory linking by Non-Orthogonal Procrustes Method. There were 3 conditions of proportion common items’ with mixed format test consisting of 20%, 25% and, 30% of total items, respectively. This research was based on data simulation using Monte Carlo Method and included 3,000 examinees. Data simulation consisted of three steps - generating true item parameters and response patterns for each grade level, calibrating multidimensional item response theory parameter model, and equating the procedure of linking. RMSE and BIAS were used as criteria to compare the quality of linking scores in each condition. Results showed that the proportion of common items affected the difference between the stability and the accuracy of linking method at .05 significant level. In other words, the difference in quality linking method depended on proportion of common items, especially the condition of 25% which approximated to multidimensional linking scores.

© 2012 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Dr. Zafer Bekirogullari of Cognitive – Counselling, Research & Conference Services C-crcs.
Keyword: Multidimensional item response theory; Mixed format test; Test equating; Anchor test; Vertical scaling

* Corresponding author. Tel.: +66-043-57-9221
E-mail address: siwa080516@hotmail.com
Introduction

Multidimensional item response theory (MIRT) model can be very helpful in trying to better understand what an item measure, what an examinee’s proficiency level is on each trait, and how accurately the different composites of ability are being assessed. Multidimensional item information and multidimensional item statistics (i.e., discrimination and difficulty) can also be use to help better understand the data structure and improve test blueprint classifications (Reckase, 2009; Yao & Boughton, 2009). A recent application of MIRT to linking for test with mixed item types by Yao & Boughton (2009) found that the equating can be performed which allows one to compare scores on subscales with fewer items across forms, while borrowing information from the other subscales to reduce estimation error. An assessment with subscale diagnostic score reporting can provide a much more detailed of where the student needs remediation and, with equating across years, student’s growth in each of the subscales can be traced. Thus, there is a need for research to develop MIRT model and linking procedures for test that comprise both multiple choice (MC) and construct response (CR) items (i.e., mixed format), because many assessment use both types. In equating process, the proportion of common item set is also crucial for successfully equating multiple mixed-format test design. The proportion of common item sets under the multidimensional test situation due to multiple item format will be of primary concern in this study. Previous research on the proportion of common item sets (Kim & Lee, 2006; Kirkpatrick, 2005; Tate, 2000) only focus on whether to include or exclude CR items in the common item sets. No research has explored the proportion of common item set representative of the total test on mixed format test equating. This research explored the effects of proportion common items’ with mixed format test on the multidimensional item response theory linking by Non-Orthogonal Procrustes Method. There were 3 conditions of proportion common items’ with mixed format test consisting of 20%, 25% and, 30% of total items. Simulated sample sizes of 3,000 examinees were use derived parameters. The condition varied in the investigation was common item set length. Root mean square error (RMSE) and bias (BIAS) were use as evaluation criteria to examine the stability and accuracy of linking.

MIRT Models and Linking Method

MIRT Models

Two types of MIRT models have been developed, compensatory (Reckase, 2009) and noncompensatory models (Simson, 1978 cited in Reckase, 2009) Since most research on MIRT has been done using compensatory models, and the fit of the two types of MIRT models appears indistinguishable from a practical point of view, the compensatory model is considered in this study.

The compensatory multidimensional extension of the three-parameter logistic model with m dimension is (M-3PL; Reckase, 2009)

\[
P(x_{ij} = 1|\theta_j, c_i, d_i, \alpha_i) = c_i + (1 - c_i) \frac{\exp(\alpha_i \theta_j + d_i)}{1 + \exp(\alpha_i \theta_j + d_i)}, \]

where \(P(x_{ij} = 1|\theta_j, c_i, d_i, \alpha_i)\) is the probability of a correct response for examinee \(j\) on test item \(i\) in an \(m\) dimension space, \(x_{ij}\)is the item response for person \(j\) on item \(i\) (1 correct; 0 wrong) \(\alpha_i\) is a vector of discrimination parameters of item \(i\), \(c_i\) is the lower asymptote (probability of correct answer when an examinee’s ability is very low), \(d_i\) is a parameter related to item difficulty of item \(i\), and is a vector of the \(j\)th examinee’s abilities.

The probability of a response of \(k\)-1 to polytomous item \(i\) for an examinee with ability vector is given by the multidimensional version of the partial credit model (M-2PPC; Yao, 2010)

\[
P_{ijk} = P(x_{ij} = k - 1|\theta_j, d_{ij}, \alpha_i) = \frac{\exp((k-1)\alpha_i \theta_j - \sum_{k=1}^{d_{ij}} d_i)}{\sum_{m=1}^{(m-1)\alpha_i \theta_j - \sum_{k=1}^{d_{ij}} d_i}}, \]

(2)
where \( P(x_{ij} = k = 1/a_i, d_{\delta k}, \theta_j) \) is the response of examinee \( j \) to item \( i \), \( d_{\delta k} \), for \( k = 1, 2, \ldots, K_i \), are the threshold parameter. \( d_{\delta k} = 0 \), and \( K \) is the number of response categories for the \( i \)th item.

The MIRT difficulty and discrimination factors are not directly equivalent to those of UIRT because of different parameterizations. Two statistics are used to capture multidimensional item characteristics corresponding to unidimensional item discrimination and difficulty. The discrimination power of a multidimensional item is (Reckase, 2009)

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

where \( MDISC_i \) denotes the \( i \)th item’s multidimensional as a function of the slope at the steepest point, and \( a_{ik} \) is the \( i \)th item’s discrimination on the \( k \)th dimension.

Multidimensional item difficulty equivalent to unidimensional difficulty is

\[
MDIFF_i = -\frac{d_i}{MDISC_i},
\]

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

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

\[
\alpha_{ik} = \arccos \frac{a_{ik}}{MDISC_i} \quad \text{(or } \cos \alpha_{ik} = \frac{a_{ik}}{MDISC_i})
\]

where \( \alpha_{ik} \) is an angle from the \( k \)th dimension.

**MIRT Linking Method**

Research on MIRT linking is a relatively new area of interest and has not been conducted as intensively as unidimensional IRT linking. However, there are some papers that describe MIRT linking methods in detail (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima et al., 1997; Yon, 2006).

Yon (2006) conducted a MIRT linking study specifically for the vertical equating situation. The NOP method suggested by Reckase & Martineau (2004) is similar to Min’s method, but the rotation matrix is non-orthogonal. By using the NOP rotation matrix, the dilation matrix is not needed as in the study by Oshima et al. (2000) is the way they estimate the rotation matrix and translation vector. Oshima et al. (2000) estimate them simultaneously, and thus has an indeterminacy problem that relates to translation, rotation, and scaling. NOP related to Yon (2006), Li & Lissitz (2000), and Min’s methods (Min, 2003) resolves the indeterminacy problem by estimating the rotation and translation matrices separately and using the procrusted linear transformation. In Li & Lissitz (2000) and Min (2003), the rotation matrix is orthogonal; that is, all items are rotated by a fixed number of degrees, thus the relationships among item dimension vectors change before and after rotation. The model for the NOP is here re-written slightly to parallel the Min (2003) model. Yon (2006) used the oblique (non-orthogonal) procrustes rotation matrix which is obtained as follows (Mulaik, 1972)

\[
T = (A'A)^{-1}A'B
\]

For an M-3PL item \( i \) let

\[
a^*_{ij} = \hat{a}_j^T T
\]

\[
d^*_i = d_i + \hat{a}_iTm
\]

For an M-2PPC item \( i \)

\[
d^*_{\delta j} = d_{\delta j} + \hat{a}_iTm \quad \text{when } k = 1, \ldots, K_i
\]

\[
\hat{\theta}_j^* = (T^{-1}\hat{\theta}_j - m)
\]
Method and Analysis

Simulation Data Analysis

Item parameter and test response patterns

Item parameters were drawn from probability distribution as to which ranges were determined by the specification of dimensional structure. One type of dimensional structures was investigated: approximate simple structure (APSS) (Roussos, Stout & Marden, 1998). For the present simulation, an APSS was constructed using two sets of items. One set of items mainly loaded on the first dimension and the other set on the second dimension. To construct dimensional structures, angles ($\alpha$) between item vectors and the first dimension were randomly drawn from a uniform distribution with given ranges of the dimensional structures.

In order to define item parameters, fix values of MDISCs and MDIFFs generated by Yon (2006) were used. The average value for MDISC of higher level is 1.0 (0.6, 0.8, 1.0, 1.2 & 1.3), and averages value of MDIFF for higher level is 0.9 (-0.5, 2.0, 1.0, 0.0 & 2.2). The average value for MDISC of lower level is 1.0 (0.6, 0.8, 1.0, 1.2 & 1.3), and averages value of MDIFF for lower level is 0.9 (0.0, 2.5, 1.5, 0.5 & 3.0). This pattern was repeated for 20 times. Discrimination and difficulty-related parameters were determined by Equation 3, 4, and 5 with given angles, MDISCs, and MDIFFs.

Finally, the probability of getting an item correct was computed by mean of a two-dimensional IRT model, and this probability was compared with a random value drawn from a uniform distribution to generate dichotomous item response (1 or 0) pattern and polytomous item response (0, 1 & 2) pattern.

Ability distribution

Three bivariate normal distributions with various mean and variances/covariances were considered for illustrating the true abilities of examinees. Different ability distributions mean that the test forms were administered to somewhat different populations. The mean and variance/covariances of group 1 is $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and group 2 is $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Number of examinees

Yao & Boughton (2007 cited in Yao & Boughton, 2009) found that a sample size of 3,000 was needed for accurate and stable parameter estimation for test that are similar to the ones used here and was therefore used for all conditions in this study.

Number of replication

Given proportion of common item (3), there were 3 combinations of simulation conditions. 10, 13, & 16 test response sets were generated for each combination.

Several computer programs were used in the simulation study. In order to generate ability distribution that was bivariate normal distribution with given mean and variances/covariances, MATLAB (Math Works, Inc, 2009) was used. For multidimensional item calibration, BMIRTII (Yao, 2010) was used. MATLAB (Math Work, Inc, 2009) were run to implement the NOP method.

Evaluation criteria and statistical tests

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

\[
BIAS = \frac{\sum_{i=1}^{n} (a_{ik} - \hat{a}_{ik})}{n},
\]

\[
RMSE = \left(\frac{\sum_{k=1}^{n} (a_{ik} - \hat{a}_{ik})^2}{n - 1}\right)^{1/2},
\]

where, $a_{ik}$ is the $i$th item parameter on the $k$th dimension, $\hat{a}_{ik}$ is the transformed value, and $n$ is the number of items.

Result
The result of the mean of root mean square error (RMSE) and bias (BIAS) were compute in Table 1. In each cell, there are seven item parameters: discrimination and difficulty of dichotomous item, discrimination and difficulty of polynomous item. The mean of RMSE and BIAS for each item parameter of linking result depended on the proportion of common item set.

Table 1. The mean of RMSE and BIAS of items parameters

<table>
<thead>
<tr>
<th>Condition</th>
<th>RMSE</th>
<th>BIAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC Item</td>
<td>CR Item</td>
</tr>
<tr>
<td></td>
<td>a₁ a₂ d₁</td>
<td>a₂ a₃ d₂ d₃</td>
</tr>
<tr>
<td>20%</td>
<td>0.765 0.618 0.498</td>
<td>0.346 0.349 0.591 1.297</td>
</tr>
<tr>
<td>25%</td>
<td>0.802 0.447 0.513</td>
<td>0.342 0.267 0.846 1.398</td>
</tr>
<tr>
<td>30%</td>
<td>0.848 0.580 0.553</td>
<td>0.247 0.300 0.972 1.342</td>
</tr>
</tbody>
</table>

The Table 1 showed that the appropriate of proportion of common item set for stability and accuracy of NOP linking method is 25%. Two summary statistics are plotted in Figure 1.

After finding significant multivariate results for the multivariate test results for seven dependent variable were compute in Table 2.
Table 2. Test statistics (F) and degree of freedom (df)=2 of RMSE and Biases from One way-MANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>RMSE</th>
<th>BIAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC Item</td>
<td>CR Item</td>
</tr>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>Proportion df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>.487</td>
<td>1.80</td>
</tr>
</tbody>
</table>

*P<.05

The results of the multivariate ANOVA showed that the proportion of common item set had non significant effects on the almost bias and RMSE of seven item parameter.

**Discussion and Conclusion**

The results from this study indicated that the appropriate of the proportion common item set for multidimensional vertical scaling on NOP method is 25%.

A frequently used equating design is common item design. Under this design, both test forms contain a subset of items that are identical across the form. The size of the subset in common is usually 20% to 25% of the total test length for unidimensional item response linking (Reckase, 2009). But, it is not clear if generalize to MIRT case. This research was found that the appropriate proportion of common item set for multidimensional linking for test with mixed item types is 25%. However, this research was studied only NOP linking method, and the approximate simple structure (APSS). For future research, it will be important to compare the approach used here to other linking procedure such as TCF method and mixed structure.

**References**


