Kasetsart Journal of Social Sciences 42 (2021) 215–220

Kasetsart Journal of Social Sciences

journal homepage: http://kjss.kasetsart.org

Kasetsart Journal of
Social Sciences

# A Comparison of the Accuracy of Multidimensional IRT equating methods for Mixed-Format Tests

Panida Panidvadtana, Siridej Sujiva*, Siwachot Srisuttiyakorn

*Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Pathumwan, Bangkok 10330, Thailand*

**Article Info**

*Article history:*
Received 3 April 2019
Revised 15 October 2019
Accepted 16 December 2019
Available online 22 February 2021

*Keywords:*
anchor test,
equating,
linking,
Multidimensionl IRT,
Mix-format test

## Abstract

This research aims to compare the accuracy of multidimensional IRT equating between concurrent calibrated MOSE procedure (CMOSE) and separated calibrated MOSE with Test characteristic function procedure (SMOSE) under the Monte Carlo simulation for mixed-format tests with approximate simple structure under the non-equivalent group with anchor test design (NEAT) with different common item score proportions. The consideration of accuracy of score equating is based on coefficient of variance of standard error of equating (CVSE). When comparing the CVSE value between the CMOSE and the SMOSE, the finding shows that there is interaction between MOSE procedures and common item proportions affecting CVSE value. The result of descriptive statistics shows that the CMOSE has lower CVSE value than that of the SMOSE. The result of simple-effect analysis shows that the CMOSE has lower CVSE value than that of the SMOSE when the proportions are 30 percent and 10 percent ($p$ = .004 and $p$ = .000, respectively). When comparing the CVSE value between the common item proportions for the CMOSE, the finding shows that the CVSE values among the proportions are not different. While the result of simple-effect analysis for the SMOSE shows that the 30% proportion has lower CVSE value than the 20% proportion ($p$ = .000), the 20% proportion has lower CVSE value than the 10% proportion ($p$ = .000) and the CVSE value of the proportion of 20 percent and 10 percent are not different. However, both types of the MOSE have the lowest CVSE when the proportions are 20 percent, 30 percent and 10 percent, respectively.

## Introduction

Score equating is the conversion of examinee score across test forms using statistical procedure to be able to compare scores between test forms accurately and fairly (Kolen & Brennan, 2004). The multidimensional IRT equating is developed to be used with the tests with more than one trait called Full MIRT equating procedure (FMIRT), consisting of 3 procedures which are (1) Full MIRT observed score equating (MOSE) procedure, (2) Unidimensional approximation to MIRT observed score equating (AOSE) procedure and (3) Unidimensional approximation to MIRT true score equating (ATSE) procedure.

According to previous research on MIRT score equating, the result shows that the MOSE is the most efficient for the mixed-format tests and dichotomous items under the random groups design (Lee, 2013; Peterson, 2014). This research adopts separate calibrations using scale linking before score equating. The previous research shows that the test characteristic function (TCF) procedure provides a good equating result (Zhang, 2012; Yao & Boughton, 2009). However, the previous research on MIRT and UIRT scale linking shows that the concurrent calibration for each testing group has lower error than separate calibration (Kim & Kolen, 2006; Lin, 2008; Simon, 2008; Tian, 2011). Furthermore,

* Corresponding author.
E-mail address: ssiridej@chula.ac.th (S. Sujiva).

https://doi.org/10.34044/j.kjss.2021.42.1.34
2542-3151/© 2021 Kasetsart University.

there has not been any study of MIRT equating with mixed-format tests under the non-equivalent groups with anchor test (NEAT) design with concurrent calibration. Therefore, the researchers deem that the development of MOSE with concurrent calibration is necessary to widen the theoretical concept of MIRT equating.

It is necessary for an efficient score equating to consider various factors such as common item score proportions and test structures. There are 3 types of test structures which are simple structure (SS), approximate simple structure (APSS) and complex structure (CS). However, the SS does not correspond with the actual condition of the test as it is difficult to have one item in only one dimension. Moreover, the test structure does not have much impact of efficiency of score equating (Lee, 2013; Peterson, 2014; Zhang, 2012) and researchers therefore choose to study the APSS for this research.

When considering the common item score proportions, Angoff et al. (1971) suggested that for the test with 40 items, the common item must be applied to at least 20 percent of all items and for the test with more than 40 items, the common item must be applied to at least 30 percent of all items. For MIRT equating, there has not been any study on an appropriate common item score proportions. The common item of 50 percent is used for MIRT equating (Peterson, 2014) which is considered high when comparing with the previous principle of common item.

For this reason, the researchers are interested in comparing the accuracy of score equating between the MOSE with concurrent calibration and MOSE with separate calibration and TCF scale linking for the APSS under NEAT design when the common item score proportions are different, namely,10 percent, 20 percent and 30 percent, to widen the concept of MOSE with appropriate resource utilization and accurate score equating results.

## Literature Review

### *Concept and Related Research on Full MIRT Observed Score Equating Procedure (MOSE)*

The MOSE procedure consists of 3 main steps as follows;

1. Calculating conditional observed score distributions in each ability level with R using the instructions of Brossman (2010). Details are as follows;

1.1 Calculating conditional observed score distributions ($f_r(\theta_j)$) for dichotomous items using Lord–Wingersky algorithm under MIRT as shown in below Equation (1).

$$\begin{aligned} f_r(\theta_j) &= f_{r-1}(\theta_j)(1-P_r) \text{ where } x = 0 \\ f_r(\theta_j) &= f_{r-1}(\theta_j)\, P_r \text{ where } x = r \\ f_r(\theta_j) &= f_{r-1}(\theta_j)(1-P_r) + f_{r-1}(\theta_j)\, P_r \text{ where } 0 < x < r \end{aligned} \tag{1}$$

where $\theta$ is ability vector and $P_r$ is the probability of correct responses for item index $r$

1.2 Calculating Conditional observed score distributions for polytomous items ($f_r(\theta_j)$) using the formula of Hanson (1994 cited in Peterson, 2014) and Thissen, Pommerich, Billeaud, and Williams (1995 cited in Peterson, 2014) as follow in Equation (2):

$$\begin{aligned} & f_r(x \mid \theta_j) = \textstyle\sum_{k=1}^{K_j} f_{r-1}(x - Wrk)\, P_{rk}(\theta_j) \\ & \text{where } min_r < x < max_r \end{aligned} \tag{2}$$

where $K_i$= the highest score for $i$, $Wrk$ is the scoring function of $k$ for $r$ item, $min_r$ is the possible lowest score after the addition of items for $r$ and $max_r$ is the possible highest score after the addition of items for $r$

1.3 Calculating conditional observed score distributions for mixed-format tests by combining distribution results of both items on the ability vector ($\theta_j$)

2. Calculating marginal observed score distributions for mixed-format tests by multiplying the conditional observed score distributions (Clause 1.3) with the multivariate ability density ($\Psi(\theta)$) and combining all results of $m$ latent ability as in the Equation (3).

$$f(x) = \textstyle\sum_{\theta 1} \sum_{\theta 2} \cdots \sum_{\theta m} f(x \mid \theta)\Psi(\theta) \tag{3}$$

3. Equating score with the traditional equipercentile method using marginal observed score distributions (Peterson, 2014)

The score equating for polytomous items under random groups shows that the MOSE is the most efficient (Lee, 2013). Furthermore, the mixed-format test score equating under random groups shows that the MOSE is more efficient than equating with UIRT and Bi-factor method (Peterson, 2014) which indicates that the MOSE is appropriate to be used for MIRT equating under NEAT. There has not been any study on mixed-format test for MIRT equating. The separate calibration is also used for previous MIRT equating research. However, according to the previous research on MIRT and UIRT scale linking, the result shows that the parameter calibration for each testing group has lower error than separate calibration (Kim & Kolen, 2006; Simon, 2008; Tian, 2011).

### *Concept and Related Research on Test Characteristic Function (TCF) Procedure*

The TCF scale linking aims to find the lowest sum of the difference between the Test characteristic surface (TCS) of multidimensional test. The unweighted TCF method is as the following Equation (4).

$$\begin{aligned} & min\left\{\tfrac{1}{Q^D}\textstyle\sum_{i=1}^{Q^D}\left[TRF(\vec{\theta}_i,\vec{\beta}_i) - TRF(\vec{\theta}_i,\vec{\beta}^*)\right]^2\right\} \\ & TRF(\vec{\theta}_i,\vec{\beta}_i) = \frac{1}{J_1+\sum_{j=1}^{J_2}(K_j-1)}\textstyle\sum_{i=1}^{N}\left[\sum_{j=1}^{J_1} P_{ij1} + \sum_{j=1}^{J_2}\sum_{k=1}^{K_j}(k-1)P_{ijk}\right] \end{aligned} \tag{4}$$

where $Q^D$ is the quadrature points, $P_{ijk}$ is the response probability in $k$–1 for the $j$ polytomous items of the $\vec{\theta}_i$ ability vector, obtained from MGPC, $D$ is the number of dimensions, $J_1$ is the number of dichotomous items and $J_2$ is the number of polytomous items (Simon, 2008)

There are MIRT scale linking methods such as Min (M)'s method, Reckase and Mastineau (NOP)'s method, equated function procedure, test characteristic function (TCF) procedure and item characteristic function (ICF) procedure etc. Th research on score equating with MIRT approach shows that the FMIRT with TCF and ICF scale linking methods shows better efficiency than the M and NOP's method (Zhang, 2012), which is similar to the research on scale linking with MIRT approach, which shows that the scale linking for multidimensional test with TCF method under NEAT design shows a good scale linking result (Yao & Boughton, 2009). Therefore, the TCF method is appropriate for scale linking between multidimensional test with separate calibration.

Some of the previous researches with MIRT equating use actual data from standardized-tests with large sample size (Brossman, 2010; Peterson, 2014), but others use simulated data because there are not real data suitable for their factor (Lee, 2013; Zhang, 2012). Standardized tests developed with MIRT model, mix-format test and common items are not available in Thailand, so the researchers needed to simulate data with the conditions. According to the previous research findings and limitations, the researchers are interested in comparing the accuracy between CMOSE and SMOSE for mixed-format tests with a simple structure under NEAT design when the common item score proportions are different with Monte Carlo simulation.

*Accuracy of Score Equating*

The accuracy of score equating is considered from coefficient of variance of standard error of equating (CVSE), which means the dispersion of standard error of equating as in equation (4). Standard error of equating is as in Equation (5).

$$CVSE = \frac{SE_i}{\overline{\hat{e}_{base}(x_i)}} \times 100 \; (11) \text{ and } SE_i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[\hat{e}_{base_k}(x_i) - \overline{\hat{e}_{base}(x_i)}]^2} \qquad (5)$$

Where $\hat{e}_{base}(x_i)$ is the equating score from the studied score equating method and $\overline{\hat{e}_{base}(x_i)}$ is the mean of equating score from the studied score equating method.

## Methodology

*The Conditional Data Model*

The conditional data model is carried out for 2 groups of 3,000 sample. Each of the mixed-format tests consist of dichotomous and polytomous items. Total score of each test is 100 and dichotomous item score ratio of each test is 70:30. The common item score proportions are 10 percent, 20 percent and 30 percent and dichotomous common item score ratio of each proportion is 70:30 too, as shown in Table 1.

*Research Conditions-Based Data Simulation*

The data simulation consists of 4 steps, which are simulation of ability of testing group, simulation of item responses, calibration, TCF scale linking for SMOSE and score equating.

*Simulation of ability of testing group*

The simulation of the ability of both testing groups uses "mvtnorm" package in R. The ability of testing group has a normal distribution. The mean vector and variance-covariance matrix of the first group is $\begin{bmatrix}0\\0\end{bmatrix}$ and $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ respectively. The mean vector and variance-covariance matrix of the second group is $\begin{bmatrix}0.5\\0.5\end{bmatrix}$ and $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ respectively. The correlation between ability dimensions is 0.5.

*Simulation of item responses*

The simulation of item responses starts from specifying (1) multidimensional difficulty (MDIFF) and overall discriminating power of an item (MDISC) in 5 levels according to the concept of Min (2003 cited in Zhang 2012). The items are divided into 5 levels. The 5 levels of MDISC are 0.4, 0.8, 1.2, 1.6 and 2.0, respectively and the 5 levels of MDIFF are -1.5, 1.0, 0, -1.0 and 1.5, respectively, and (2) angles in each dimension in which the items are divided into 2 subgroups. The vector of item in the first dimension angles with dimension 1 ($\theta_1$) at 0–15 degrees. The vector of item in the second dimension angles with dimension 1 ($\theta_2$) at 76–90 degrees. The angle of each item in each subgroup is selected randomly by uniform distribution according to the range and number of items specified.

Afterwards, the MDIFF, MDISC and vector angles are used to measure the parameter of the item as shown in Equation 4 when $MDIFF_i$ is MDIFF for *i* item, $MDISC_i$ is MDISC for *i* item, $a_{1i}$ and $a_{2i}$ are vectors of discriminating power of parameter of the first and second dimension for *i* item respectively, $d_i$ is scalar parameter, which relates to the difficulty of the test for *i* item, and $a_{1i}$ and $a_{2i}$ are vector angles of the first and second dimension for *i* item (Equation (6)).

$$a_{1i} = MDISC_i \times cosa_{1i},\ a_{2i} = MDISC_i \times cina_{1i} = MDISC_i \times cosa_{2i} \text{ and } d_i = MDISC_i \times MDISFF_i \qquad (6)$$

Then, calculate the probability of correct responses for polytomous items with multidimensional extension of the tree-parameter logistic (M3PL) and the probability of correct responses for dichotomous items with Multidimensional generalized partial credit model with 2 parameters (MGPC), as shown in Equation 5 and 6, respectively, where $\theta_j$ = ability vector of *j* testing person, $K_j$ = highest score for *i* item i where the score range is 0, 1 and 2 and $\beta_{iu}$ is the threshold parameter for *u* score.

**Table 1** The dichotomous common item score ratios with different common item score proportions

| The common item score proportions | Score (items) | | |
|---|---|---|---|
| | Dichotomous | Polytomous | Total |
| 10% (10 score) | 8 (8) | 2 (1) | 10 (9) |
| 20% (20 score) | 14 (14) | 6 (3) | 20 (17) |
| 30% (30 score) | 22 (22) | 8 (4) | 30 (26) |

$$\text{M3PL: } P(\theta_j, a_i, c_i, d_i) = c_i + (1 - c_i)\frac{e^{a_i\theta_j' + d_i}}{1+e^{a_i\theta_j' + d_i}} \tag{7}$$

$$\text{MGPC: } P(u_{ij} = \mathrm{k}|\theta_j) = \frac{e^{k\mathrm{a}_i\theta_j' + \sum_{u=o}^{k}\beta iu}}{\sum_{u=o}^{k_i} e^{v\mathrm{a}_i\theta' + \sum_{u=o}^{k}\beta iu}} \tag{8}$$

Then, use the probability of correct responses to simulate item responses with uniform random. The simulation of responses of polytomous items for *i* item for *j* testing person with the *θ* ability ($U_{ij}$) is as follows; $U_{ij} = 0$, $y > P_{ij}$ and $U_{ij} = 1$, $y \leq P_{ij}$. The simulation of responses of dichotomous items for *i* item in the k response list for the *j* testing person with the *θ* ability ($U_{ijk}$) is as follows; $U_{ijk} = 0$, $0 < y \leq P_{ij1}$, $U_{ijk} = 1$, $P_{ij1} < y \leq P_{ij2}$ and $U_{ijk} = 2$, $P_{ij2} < \mathrm{y} < 1$

*MIRT Calibration*

The calibration for multidimensional mixed-format test uses M3PL and MGPC using "mirt" package in R with EM algorithm. The separate calibration uses mirt () and the concurrent calibration uses multiple Group ().

*Scale linking*

The scale linking uses the unweighted test characteristic function (TCF) as shown in Equation 3 by specifying the 2×2 rotation matrix and 2×1 translation matrix with plink package in R in accordance with Oshima, Davey and Lee (2000, cited in Zhang, 2012) with a non-orthogonal rotation. The obtained matrix is then used to convert the parameter.

*Score equating*

The MOSE score equating in R consists of 3 steps which are (1) calculating the conditional observed score distributions, (2) calculating the marginal observed score distributions (3) traditional equipercentile method using marginal observed score distributions through the "equate" package. The details are mentioned above. The simulation of multivariate ability density ($\Psi(\theta)$) of *m* latent ability adopts the multivariate standard normal distribution with uncorrelated axes ($\theta \sim MVN(0,I)$ which is measured from the "mvtnorm" package in R.

*Analysis of Accuracy of Score Equating*

The accuracy of score equating is considered from the coefficient of variance of standard error of equating (CVSE). The comparison of the efficiency of both types of MOSE when the common item score proportions are different uses 2 types of analytic statistics, which are descriptive statistics and two-way ANOVA

## Results

*Descriptive Statistics of Efficiency of Score Equating*

When considering the coefficient of variation of standard error (CVSE) of both types of MOSE, the result shows that the CMOSE has lower CVSE value than SMOSE in every condition. When considering the CVSE value of common item score proportions, the result shows that both types of MOSE have the lowest CVSE value when the common item score proportions are 20 percent, 30 percent and 10 percent, respectively in every condition as shown in Table 1.

*The Comparison of the Efficiency of Score Equating Between Both Types of MOSE*

The analysis result of two-way ANOVA shows that there is interaction between MOSE procedures and common item score proportions, which affects the CVSE value, with a statistical significance of .05 (*p* = .006). Therefore, the simple-effect analysis is adopted.

The simple-effect analysis result for both types of MOSE procedures when the proportions are 30 percent and 10 percent shows that the CMOSE has lower CVSE value than the SMOSE with a statistical significance of .05 (*p* = .004 and *p* = .000, respectively), whereas, when the proportion is 20 percent, the result shows that the CMOSE has lower CVSE value than the SMOSE with no statistical significance of .05 (*p* = .006)

The simple-effect analysis result for common item score proportions of CMOSE shows that each common item score proportion does not have a statistically-significant difference of .05. For SMOSE, the result shows that the 30% proportion has lower CVSE value than the 10% proportion with the statistical significance of .05 (*p* = .000) and the 20% proportion has lower CVSE value than the 10% proportion with the statistical significance of .05 (*p* = .000) whereas the CVSE value of the 20% and 10% proportions are not different (*p* = .254) as shown in Table 2.

## Discussion

The comparison of the accuracy of score equating between MOSE when the common item score proportions are 30% and 10% shows that the CMOSE has lower CVSE value than the SMOSE with the statistical significance of .05, but when the common item score proportion is 20%, the result shows that the CMOSE has lower CVSE value than the SMOSE with no statistical significance of .05. However, the descriptive statistics show that the CMOSE has lower CVSE value in every condition, indicating that the CMOSE has more accuracy of score equating than the SMOSE as the one-time calibration

**Table 1** Mean and standard deviation of CVSE value

| Common item score proportions (%) | CVSE value of CMOSE | | CVSE value of SMOSE | |
|---|---|---|---|---|
| | Mean | *SD* | Mean | *SD* |
| 30 | 7.627 | 10.969 | 12.4191 | 10.792 |
| 20 | 5.905 | 6.205 | 10.508 | 8.268 |
| 10 | 8.554 | 9.407 | 19.874 | 20.428 |

**Table 2** Simple-effect analysis result affecting CVSE value

| | | | Mean Difference (I-J) | *p* | 95% Confidence Interval for Difference | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Comparing CVSE value of MOSE in each common item score proportion | | | | | | |
| 30% | CMOSE | SMOSE | -4.792 | .004* | -8.080 | -1.504 |
| 20% | CMOSE | SMOSE | -4.603 | .006 | -7.892 | -1.315 |
| 10% | CMOSE | SMOSE | -11.320 | .000* | -14.608 | -8.301 |
| MOSE Comparing CVSE value of common item score proportion in each MOSE procedure | | | | | | |
| CMOSE | 30% | 20% | 1.722 | .304 | -1.566 | 5.011 |
| | 30% | 10% | -0.927 | .580 | -4.215 | 2.361 |
| | 20% | 10% | -2.649 | .114 | -5.938 | 0.639 |
| SMOSE | 30% | 20% | 1.911 | .254 | -1.377 | 5.199 |
| | 30% | 10% | -7.455$^*$ | .000* | -10.743 | -4.166 |
| | 20% | 10% | -9.366$^*$ | .000* | -12.654 | -6.077 |

for two tests results in a large calibration sample causes the error in calibration to be low and can reduce the error in score equating due to scale linking whereas, it is necessary for the separate calibration to use scale linking to place the obtained parameter value on the scale on the common item, which causes a random error due to the random sampling and estimation of coefficient of scale linking (Kim & Kolen, 2006; Meng, 2007). This is in line with the research on MIRT scale linking, which shows that the concurrent calibration has lower mean square error than separate calibration and uses scale linking when common item score proportion is 33.33 percent (Kim & Kolen, 2006). Moreover, the concurrent calibration is more efficient than calibration and separate calibration (Lin, 2008; Simon, 2008)

The comparison of the accuracy of score equating between the common item score proportions for the CMOSE shows that each common item score proportion does not show a statistically-significant difference of .05. The simple-effect analysis result for SMOSE shows that the 30% proportion has lower CVSE value than the 10% proportion with a statistical significance of .05, and the 20% proportion has lower CVSE value than the 10% proportion with a statistical significance of .05. However, the descriptive statistics show that both MOSE procedures have the lowest mean CVSE when the proportions are 20 percent, 30 percent and 10 percent, respectively, indicating that the proportion of 10 percent should be applied for equating.

Some of the subgroups for the 10% proportion only have one item, which causes such item to not represent the content and characteristics of statistics of the entire test. The principle of creating a common item is the common item must have an appropriate length and can represent the content and difficulty of the entire test (Kolen & Brennan, 2004). Angoff et al. (1971) suggested that for the test with 40 items, the common item must be applied to at least 20 percent of all items and for the test with more than 40 items, the common item must be applied to at least 30 percent of all items. This is in line with the research which shows that the concurrent calibration when the common item score proportion is 20 percent is more efficient than the proportion of 10 percent since when the proportion increases, RMSE, absolute bias and standard error of equating decrease (Meng, 2007)

## Conclusion and Recommendation

The research findings can be concluded into 3 aspects. First, there is the interaction between MOSE procedures and common item score proportions which affects the CVSE value. Second, the CMOSE has lower CVSE value than that of the SMOSE with the statistical significance of .05 when the common item ratios are 30 percent and 10 percent. However, the CMOSE has lower CVSE value than that of the SMOSE in every condition, which indicates that the CMOSE is more accurate than the SMOSE. Therefore, the CMOSE should be applied for MIRT equating.

Third, The CVSE value among the common item proportions for CMOSE are not different. The result for SMOSE shows that the 30% proportion has lower CVSE value than the 10% proportion, and the 20% proportion has lower CVSE value than the 10% proportion. However, the descriptive statistics shows that both types of MOSE procedures for SMOSE have the lowest mean CVSE when the common item proportions are 20 percent, 30 percent and 10 percent, respectively. Thus, the proportions of 20 percent and 30 percent should be used when equating score with both types of MOSE.

The recommendations for research in the future are as follow. First, the study should be conducted on the non-compensatory MIRT model. Second, other MIRT scale linking methods, such as, Item Characteristic Function (ICF), Direct method (OD) and Min's method, may be used. Third, an analysis program for concurrent calibration taking shorter time than R, such as flex MIRT, may be used. And, forth, the study should evaluate precision of equating.

## Conflict of Interest

There is no conflict of interest

## Acknowledgments

The research was funded by the 90th anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund).

## References

Angoff, W. H., Thorndike, R. L., & Lindquist, E. F. (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Brossman, B. G. (2010). *Observed score and true score equating procedures for multidimensional item response theory.* (Unpublished doctoral dissertation). University of Iowa, Iowa, IA.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items.* (Unpublished research note). Iowa City, IA: ACT, Inc.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357–381.

Kolen, M. J., & Brennan, R. L. (2004). Test Equating, scaling, and linking: Methods and practices. In S. E. Fienberg & W. J. v. d. Linden (Eds.), *Statistics in social sciences and public policy*. New York, NY: Springer Science Business Media.

Lee, E. (2013). *Equating multidimensional tests under a random groups design: A comparison of various equating procedures* (Unpublished doctoral dissertation). University of Iowa, Iowa, IA.

Lin, P. (2008). *IRT vs. Factor analysis approaches in analyzing multigroup multidimensional binary data: The effect of structural orthogonality, and the equivalence in test* (Unpublished doctoral dissertation). University of Maryland, Maryland, MD.

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling* (Unpublished doctoral dissertation). University of Iowa, Iowa, IA.

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

Peterson, J. L. (2014). *Multidimensional item response theory observed score equating methods for mixed-format tests* (Unpublished doctoral dissertation). University of Iowa, Iowa, IA.

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

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration mixed-format tests in the non-equivalent group common-item design under IRT* (Unpublished doctoral dissertation). Boston College University, Massachusetts, MA.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39–49.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, *46*(2), 177–197. doi: 10.1111/j.1745-3984.2009.00076.x

Zhang, O. (2012). *Observed score and true score equating for multidimensional item response theory under nonequivalent group anchor test design* (Unpublished doctoral dissertation). University of Florida, Florida, FL.