Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 40

Equating Multidimensional Tests under a Random Groups Design: A Comparison of Various Equating Procedures

Eunjung Lee[†]

Won-Chan Lee

Robert L. Brennan

November 2014

[†] Eunjung Lee, 10A Smythe Ave, Mont Albert, VIC, 3127 Australia (email: ejlee79@gmail.com). Won-Chan Lee is Associate Professor and Co-director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu). Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Co-director, CASMA, 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

The purpose of this research was to compare the performance of various equating procedures for multidimensional tests. To examine the various equating procedures, simulated data sets were used that were generated based on a multidimensional item response theory (MIRT) framework. Specifically, the performance of the following six equating procedures under the random groups design was compared: (1) unidimensional IRT observed score equating, (2) unidimensional IRT true score equating, (3) full MIRT observed score equating, (4) unidimensionalized MIRT observed score equating, (5) unidimensionalized MIRT true score equating, and (6) equipercentile equating. A total of four factors (test length, sample size, form difficulty differences, and correlations between dimensions) were expected to impact equating performance, and their impacts were investigated by creating two conditions per each factor: long vs. short test; large vs. small sample size; some vs. no form differences; and high vs. low correlation between dimensions. To evaluate the performance of each equating procedure, the population equating relationships are defined in two ways: the FMIRT result and the identity equating. The conditional and overall equating relationships are evaluated in terms of bias, standard error, and overall error.

The following findings are notable: (1) the full MIRT procedure provided more accurate equating results than other equating procedures especially when the correlation between dimensions was low; (2) the equipercentile procedure was more likely than the IRT methods to yield a larger amount of random error and overall error across all the conditions; (3) equating for multidimensional tests was more accurate when form differences were small, sample size was large, and test length was long; (4) even when multidimensional tests were used (i.e., the unidimensionality assumptions were violated), the unidimensional IRT procedures were found to yield quite accurate equating results; and (5) whether an observed or a true score equating procedure was used did not seem to yield any differences in equating results.

1 Introduction

Often multiple traits are required to successfully solve a problem in many educational or psychological tests. Accordingly MIRT models have been receiving greater attention in recent years, and have been adopted to analyze those tests psychometric properties (Reckase, 2009). However, when MIRT models are used, scale linking and equating become much more complex. Little research exists in the literature on MIRT equating, and the present study is intended to help fill the gap by conducting a comprehensive simulation study to evaluate the performance of various equating methods for multidimensional data.

The primary objective of this study is to compare the equating performance of the various equating procedures for the multidimensional tests using simulated data sets. The equating procedures include both unidimensional and multidimensional equating procedures based on an IRT framework as well as a traditional equipercentile equating procedure. Specifically, six equating procedures under the random groups design were compared: (1) UIRT observed score equating (UIRTO), (2) UIRT true score equating (UIRTT), (3) full MIRT observed score equating (FMIRT), (4) unidimensionalized MIRT observed score equating (UMIRTO), (5) unidimensionalized MIRT true score equating (UMIRTT), and (6) unsmoothed equipercentile equating (EQ). Factors investigated include test length, sample size, form difficulty differences, and correlations between dimensions.

2 Theoretical Framework

This chapter provides a brief review of unidimensional IRT (UIRT) equating procedures and MIRT equating procedures.

2.1 Unidimensional IRT Equating

For tests that are intended to measure one dimension, a unidimensional IRT equating procedure can be used. A UIRT equating procedure can be used with different types of equating designs. Because the random groups equating design is used in this study, the focus of the following description is on IRT equating procedures in conjunction with the random groups design. If a specific IRT model fits a set of data, the IRT parameters for each form are estimated separately in the random groups equating design. In the random groups equating design, the separately estimated parameters for the two forms are assumed to be on the same scale without further transformation when ability distributions are specified to be the same in both calibrations (Kolen & Brennan, 2004).

The steps for IRT true score equating are as follows. First, a true score on the new form is selected. Second, a specific θ that corresponds to the new-form true score is found using an iterative procedure, such as the Newton-Raphson method. Third, the true score on the old form that corresponds to the specific

θ is obtained using the old-form test characteristic curve (Kolen & Brennan, 2004).

To conduct IRT observed score equating, the observed score distribution for examinees with a given θ is first estimated using a recursion formula (Lord & Wingersky, 1984). The observed score distributions for examinees with various θs are then found by multiplying the ability density and integrating (or summing) across all the θs. Finally, in order to determine the equating relationship between the two forms, equipercentile equating is conducted using the marginal observed score distribution for each form (Kolen & Brennan, 2004).

2.2 Multidimensional IRT Equating

To analyze a data set for equating using a MIRT framework, item parameters are first separately estimated for each form. When a random groups design is used, under the UIRT framework, the parameter estimates do not need to be transformed. Unlike the UIRT models, however, using a random groups design does not yield the parameter estimates on the same scale under the MIRT framework. The parameters are still subject to rotational indeterminacy (Thompson, Nering, & Davey, 1997). Several researchers (e.g., Thompson et al., 1997; Yon, 2006) have developed procedures to link scales within the MIRT framework, and one of these procedures can be used to link scales for old and new forms. However, Brossman and Lee noted that the MIRT scale linking procedures under the random groups design may not always be a prerequisite to conduct MIRT equating. They are not required under certain conditions, including (1) the specified variance-covariance matrix for the ability estimates is the identity matrix, and (2) item parameters and ability estimates are calibrated with respect to orthogonal reference axes (Brossman, 2010; Brossman & Lee, 2013).

Once the linking procedure is conducted appropriately when it is required, an equating procedure is then conducted to relate scores on the two forms to be equated. It should be noted, however, that very little research has been conducted on equating with multidimensional data in the MIRT framework. Three studies were conducted by Brossman and Lee (Brossman, 2010; Brossman & Lee, 2013; Lee & Brossman; 2012). Lee and Brossman (2012) developed an observed score equating procedure using a simple structure MIRT framework. This procedure can be used for a test that consists of different (yet correlated) clusters of items; each cluster measures a different latent ability, and each item within a cluster measures the same unidimensional latent ability. In the other two studies using real data sets, they (Brossman, 2010; Brossman & Lee, 2013) developed three equating procedures (two observed score equating procedures and one true score equating procedure) in the MIRT framework and then compared the performance of the three procedures with the UIRT and the equipercentile equating procedures. The three equating procedures developed in Brossman (2010) and Brossman and Lee (2013) were based on a complex MIRT framework for each item within a cluster measures more than one latent ability dimension (i.e., each item measures multidimensional proficiencies). The

focus of this paper is on the complex MIRT framework and accordingly the three MIRT procedures by Brossman (2010) and Brossman and Lee (2013) are briefly described below

2.2.1 Full MIRT observed score equating procedure

Brossman and Lee's (Brossman, 2010; Brossman & Lee, 2013) full MIRT observed score equating procedure (FMIRT) is a relatively straightforward extension of UIRT observed score equating. Similar to the UIRT framework, the Lord-Wingersky algorithm, using a vector of ability levels in place of a single ability level, can be used in order to compute the conditional observed score distributions for each combination of ability levels in the entire ability space. In the UIRT framework, a marginal observed score distribution for each form is obtained by multiplying these conditional distributions by the ability density and then by either summing or integrating over all ability levels. The MIRT analog is the same except for using the multivariate ability density to multiply the conditional distributions. Similar to the unidimensional case, equipercentile equating is used to equate the two forms.

2.2.2 Unidimensionalized MIRT equating procedure

Brossman (2010) and Brossman and Lee (2013) developed both the observed score and true score unidimensional approximation of the MIRT equating procedures by estimating unidimensional item parameters and unidimensional ability distributions from the multidimensional data using the work by Zhang (1996), Zhang and Stout (1999), and Zhang and Wang (1998). According to these authors, a UIRT model with estimated unidimensional ability and item parameters can be used as a close approximation of a multidimensional compensatory IRT model that adequately models a set of item responses. Using their procedures, the following unidimensional parameters can be obtained from multidimensional parameters: (1) a unidimensional composite ability, which is a linear composite of multidimensional abilities, and (2) unidimensional item parameters, which correspond to the unidimensional composite ability. Using estimates of these unidimensional parameters, the standard UIRT observed score equating and true score equating procedures can be performed. For the unidimensional approximation of MIRT true score equating, the composite true score associated with the composite ability level is the sum of the probabilities of obtaining correct responses over all items at each composite ability level. For the unidimensional approximation of MIRT observed score equating, the conditional distributions can be determined at each composite ability level. In this paper, the unidimensional approximations of MIRT equating procedures are called the unidimensionalized MIRT equating procedures (UMIRT) to better represent the aforementioned nature of their equating processes.

2.2.3 Summary statistics for the MIRT model

Reckase and McKinley (1991) developed two summary statistics for the compensatory MIRT model: MDISC and MDIFF. These statistics are analogous to the discrimination and difficulty parameters in the UIRT model, respectively. These discrimination and difficulty-related parameters are:

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

and

$$MDIFF_i = \frac{-d_i}{MDISC_i}. \tag{2}$$

In Equations (1) and (2), $MDISC_i$ denotes the i-th item's discrimination; m is the number of dimensions in the ability space; a_{ik} is the i-th item's discrimination on the k-th dimension; d_i is the item intercept parameter; and $MDIFF_i$ is the distance between the origin and the steepest point of the item response surface. *MDIFF* is computed using the standard distance formula, and a difficult item has a high positive *MDIFF* value.

The $MDISC_i$ is directly related to the angle between each coordinate axis and the point of the steepest slope. These angles are determined by computing the direction cosine corresponding to each dimension:

$$cos\gamma_{ik} = \frac{a_{ik}}{MDISC_i}, \tag{3}$$

where γ_{ik} denotes the angle between the k-th axis and the line from the origin to the point in radians and the expression *cos* refers to cosine.

3 Method

This chapter consists of five main sections. First, a description of the equating design and the equating procedures used in this study is provided. Second, a description of the simulation factors examined in this study is presented. Third, a description of how the item parameters are generated is given. Fourth, a description of item calibration and equating procedures is presented. Lastly, a framework for evaluating the equating procedures is presented.

3.1 Equating Design and Equating Procedures

The random groups equating design was used in this study. In the random groups equating design, because test forms are randomly assigned to examinees, it is thus assumed that the ability distribution is the same across groups and the differences on scores are attributed solely to test form differences. Each of the following six procedures was conducted to equate scores on Form X to scores on Form Y: (1) UIRT observed score equating (UIRTO), (2) UIRT true

score equating (UIRTT), (3) full MIRT observed score equating (FMIRT), (4) unidimensionalized MIRT observed score equating (UMIRTO), (5) unidimensionalized MIRT true score equating (UMIRTT), and (6) equipercentile equating (EQ).

3.2 Factors Studied

The current simulation study included the following four factors: test length, sample size, correlation between dimensions, and form differences. Only a two-dimensional structure was considered in this study. The ability vector was modeled as bivariate normal with a mean of zero and a standard deviation of one for both dimensions. The correlations between the two dimensions were varied as indicated later.

Two examinee sample sizes were employed: 2,000 and 6,000. There were two conditions for test length: a 30-item test and a 60-item test. The test lengths for Forms X and Y for both the 30- and 60-item tests were set to be equal. There were two conditions for the correlation between two dimensions: 0.5 and 0.8. Note that, the degree of correlation between two dimensions does not differ across groups who take each form of the test, because in the random groups equating design, it is assumed that all the examinees are from the same population.

The current study included two conditions for the form differences: identical forms (IF) vs. different forms (DF). In the IF condition, no form differences were assumed in order to examine how each equating procedure performs in an ideal situation where no equating is required (i.e., Forms X and Y are identical). In the DF condition, where some degree of form differences existed, it was assumed that the forms differed in their difficulty levels. There were two conditions for the DF case: small and large differences in difficulty levels. In both conditions, item parameters for Form X were determined to have the same values of $MDISC_i$ as the corresponding Form Y but to have a higher $MDIFF_i$ mean. Specifically, the $MDIFF_i$ mean for Form X was set to be about 0.05 and 0.2 higher than for Form Y in the small and large difference conditions, respectively, which made Form Y easier than the corresponding Form X. Form X with the small form difference conditions is referred to as Form XS, and Form X with the large form difference conditions is referred to as Form XL, hereafter.

The simulation conditions for this study were not fully crossed. When no form differences were assumed, there were eight combinations of simulation conditions with two sample sizes, two test lengths, and two correlations between dimensions. On the other hand, when some degree of form differences was assumed, there were 16 combinations of simulation conditions with two sample sizes, two test lengths, two correlations between dimensions, and two form difficulty differences. A total of 24 conditions were simulated in this study. Simulations were replicated only 50 times for each condition because of the long estimation time for MIRT calibration.

3.3 Generation of Item Parameters

A dimensional structure, termed approximate simple structure (APSS) was considered for the current study. APSS was constructed by two sets of items. For APSS, one set of items loaded primarily on the one dimension (Dimension 1) and the other set of items loaded primarily on the other dimension (Dimension 2). APSS was manipulated by modifying the angular distance of items in the two dimensional space, which was represented by the item parameters for each form.

3.3.1 Item Parameters for Form Y (M2PL Model)

In order to define item parameters for Form Y, the fixed values of *MDISC* and *MDIFF* generated by Roussos, Stout, and Marden (1998), given in Table 1, were modified and used in this study. According to Roussos et al. (1998), these five pairs of MIRT characteristics are selected because they are realistic and cover item features that are usually found on a test.

Table 1: Five MIRT Discrimination and Difficulty Pairs in Roussos et al. (1998)

Level	MDISC	MDIFF
1	0.4	-10.5
2	0.8	1.0
3	1.2	0.0
4	1.6	-1.0
5	2.0	10.5
Mean	1.2	0.0

In this paper, the $MDIFF_i$ parameters for Form Y were randomly generated from a uniform distribution with the range formed using the midpoints of the *MDIFF* values in Table 1. The parameter for each item was directly determined from the five fixed $MDISC_i$ parameters in Roussos et al. (1998) as well as the randomly generated $MDIFF_i$ parameters using Equation (2).

For the M2PL model, APSS was assumed. For APSS, one set of items loaded primarily on one dimension (Dimension 1) and the other set of items loaded primarily on the other dimension (Dimension 2). APSS was manipulated by specifying a_{ik} parameters using fixed $MDISC_i$ parameters and randomly generated γ_{ik} values.

The a_{ik} parameters depend not only on the $MDISC_i$ values but also on the γ_{ik} values. To specify the a_{ik} parameters, in this study the items corresponding to their primary dimension were allowed to randomly fall within 15° of the first axis (the coordinate axis for the primary dimension), which was done

by randomly generating the γ_{ik} values for items corresponding to a particular dimension from a uniform distribution with the range $[0, 15]$.

3.3.2 Item Parameters for Form X (M2PL Model)

For the DF condition, where some degree of form differences exists, it was assumed that the forms differed in their difficulty levels. Thus, item parameters for Form X were determined to have the same values of $MDISC_i$ as the corresponding Form Y but to have a higher mean of $MDIFF_i$ (about 0.05 and 0.2 higher than that for Form Y in the small and large difference conditions, respectively). Although the items for Form X had the same $MDISC_i$ values as Form Y, the values for the two forms would not necessarily be the same because the γ_{ik} values for items corresponding to a particular dimension were randomly generated from a uniform distribution with the range $[0, 15]$.

The mean and the range of $MDIFF_i$ for Form X were manipulated by adding 0.05 to each end point of the Form Y $MDIFF_i$ range for the small form difference condition, and 0.2 for the large form difference condition. Thus, the standard deviations were intended to stay similar but the mean and the range were intended to be changed. In order to let each item have different d_i parameter values, the $MDIFF_i$ parameters for Form X were randomly generated from a uniform distribution with the *MDIFF* range for Form X. The d_i parameter for each item was directly determined from the fixed five $MDISC_i$ parameters in Roussos et al. (1998) as well as the randomly generated $MDIFF_i$ parameters using Equation (2).

3.4 Item Calibration and Equating Procedure

After responses were generated for each examinee, the item parameters for each form were estimated separately for each of the 24 conditions. Specifically, the item parameters for all the IRT models (U2PL and M2PL) were estimated using flexMIRT (Cai, 2012). flexMIRT can implement both the maximum likelihood and MH-RM (Cai, 2010) methods for item parameter estimation. Due to the limit that the number of dimensions was restricted to two in this paper, the method of maximum likelihood was used for item calibration to reduce estimation time. By using the same software for estimating the item parameters for both the UIRT model and MIRT model, a more accurate comparison between the two procedures can be performed.

When a random groups equating design is used with the UIRT equating procedures, separately estimated parameters for the two forms are assumed to be on the same scale without further transformation when ability distributions are specified to be the same in both calibrations (Kolen & Brennan, 2004). Therefore, no scale linking procedures were required for the unidimensional procedures. The estimated item parameters were used to conduct both UIRTO and UIRTT procedures using *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009).

MIRT observed score and true score equating does not require MIRT scale linking procedures under the random groups design (Brossman & Lee, 2013), if one of the following conditions is satisfied: (a) the solutions are orthogonal when calibrating item parameters and ability estimates; or (b) the ability distribution is specified to follow a multivariate standard normal distribution with zero correlation between dimensions (Brossman & Lee, 2013, pp. 470-471). Although the ability distribution in this paper was specified to follow a multivariate standard normal distribution with nonzero correlation, it turned out that the MIRT scale linking procedures were not required.

To conduct the FMIRT procedure, the computer programs R (R Development Core Team, 2008) and *Equating Recipes* (Brennan et al., 2009) were used. Specifically, R code created by Brossman (2010) was modified and then used to determine the conditional observed score distributions and the marginal observed score distributions for each form. To conduct the UMIRTO and UMIRTT, the unidimensional item parameters and the unidimensional ability distributions were estimated following the procedures of Brossman (2010) and Brossman and Lee (2013). Then, these estimates were incorporated in *Equating Recipes* (Brennan et al., 2009) to conduct both the true and observed score equating.

3.5 Evaluation Criteria

To evaluate the performance of each equating procedure, the population equating relationship should be determined. In this study, the population equating relationships were defined in two ways: the FMIRT result and the identity equating. When some degree of form differences was assumed, the population equating relationships were established by conducting FMIRT based on the item parameters (not estimates) and a population bivariate normal distribution with correlation between dimensions of 0.5 or 0.8 depending on the conditions.

Identity equating was also used in this study as an additional criterion. In identity equating, a score on Form X is considered to be equivalent to the identical score on Form Y. The Form Y equivalent of a Form X score is set equal to the Form X score (Kolen & Brennan, 2004). Identity equating reflects the ideal situation where no equating is necessary and where a score conversion yielded by any equating method equals the identity function.

3.5.1 Conditional Evaluation

Three statistics were used to evaluate the equating relationships at each of the equated score points: standard error of equating (SEE), signed bias (SB), and root mean squared error (RMSE). Each statistic was standardized by dividing each by the standard deviation of the Form Y scores (σ_Y) which was computed based on the population equating relationships.

Let x_i be a score on Form X and $\hat{t}_{yr}(x_i)$ be the equated score at x_i using an equating procedure from replication r. Then the mean of the estimates over replications is

$$\bar{\hat{t}}(x_i) = \frac{1}{R}\sum_{r=1}^{R}\hat{t}_{yr}(x_i), \tag{4}$$

where R is the number of replications (i.e., 50). The SEE for the equivalent is defined as

$$SEE(x_i) = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{t}_{yr}(x_i) - \bar{\hat{t}}(x_i))^2}}{\sigma(Y)}. \tag{5}$$

Let $t_y(x_i)$ be the equivalents of x_i based on the population equating relationship. Then the SB of the equated score is

$$SB(x_i) = \frac{\bar{\hat{t}}(x_i) - t_y(x_i)}{\sigma(Y)}. \tag{6}$$

Absolute bias (AB) is defined as

$$AB(x_i) = \frac{|\bar{\hat{t}}(x_i) - t_i(x_i)|}{\sigma(Y)}. \tag{7}$$

RMSE of the equivalent can be expressed as the square root of the sum of error variance and squared bias:

$$RMSE(x_i) = \frac{\sqrt{SEE^2(x_i) + SB^2(x_i)}}{\sigma(Y)}. \tag{8}$$

3.5.2 Overall Evaluation

Three statistics were used to estimate how well each procedure performed by aggregating errors across all score points: weighted standard error (WSE), weighted absolute bias (WAB), and weighted root mean squared error (WRMSE). For the overall statistics, a marginal observed score distribution for Form Y based on the item parameters and population distribution was used as the weight (w_i).

Letting N be the total number of items in the test, weighted standard error (WSE) is defined as:

$$WSE = \sqrt{\sum_{i=1}^{N} w_i \cdot SEE^2(x_i)}. \tag{9}$$

Weighted absolute bias (WAB) is

$$WAB = \sum_{i=1}^{N} w_i \cdot AB(x_i). \tag{10}$$

Weighted root mean squared error (WRMSE) is

$$WRMSE = \sqrt{\sum_{i=1}^{N} w_i \cdot RMSE^2(x_i)}. \quad (11)$$

4 Results

This chapter describes the conditional equating results, followed by the overall evaluation of the equating results for all the methods.

4.1 Conditional Results

Selected results are shown in Figures 1 through 3. The four terms in the parentheses of the title of each plot in Figures 1 through 3 represent the four study factors: sample size, test length, correlation between dimensions, and form differences, respectively. For example, the title of the top left plot in Figure 1 is SB(2000_30_5XL). The terms in the parentheses of the title indicate that this plot represents SB when sample size was 2,000, test length was 30, correlation between dimensions was 0.5, the form differences were large. Note that the vertical axes of plots in Figures 1 through 3 have different range.

The first notable finding was that all the equating procedures performed reasonably well throughout the whole score scale in terms of bias (i.e., SB) across all the conditions (see Figure 1). Even though all the equating procedures performed quite well, the equating trends for the two UIRT and two UMIRT equating procedures (i.e., UIRTT, UIRTO, UMIRTT, and UMIRTO) were similar to each other but different from those for the FMIRT procedure or equipercentile equating procedure (EQ). In particular, the equating trend for EQ showed the most different pattern. This pattern of findings was more apparent when sample size was small rather than large (i.e., 2,000 rather than 6,000).

Among the five IRT-based equating procedures, when some degree of form differences was assumed (i.e., XL and XS conditions), the SB lines for the two UIRT procedures (UIRTT and UIRTO) and the two UMIRT procedures (UMIRTT and UMIRTO) seemed to be close to each other, whereas the line for the FMIRT procedure was a bit apart from the other four lines. However, a slightly different pattern was found when no form difference was assumed (i.e., IF condition). That is, the two SB lines for UMIRT procedures seemed to go together, and the other three lines for the two UIRT procedures and the FMIRT procedure seemed to go together. Thus, the results showed different patterns depending on whether form differences were assumed, and the differences in results are likely to be due to the different definitions of the population equating relationship. As noted above, the population equating relationship was defined by FMIRT equating when some degree of form differences was assumed, whereas it was defined by identity equating when no form differences were assumed. Consistent with the existing findings (Brossman, 2010; Brossman &

Lee, 2013), whether an equating procedure is an observed score procedure or a true score procedure did not seem to matter much. The two UIRT true- and observed-score procedures performed similarly, and the two UMIRT true- and observed-score procedures performed similarly.

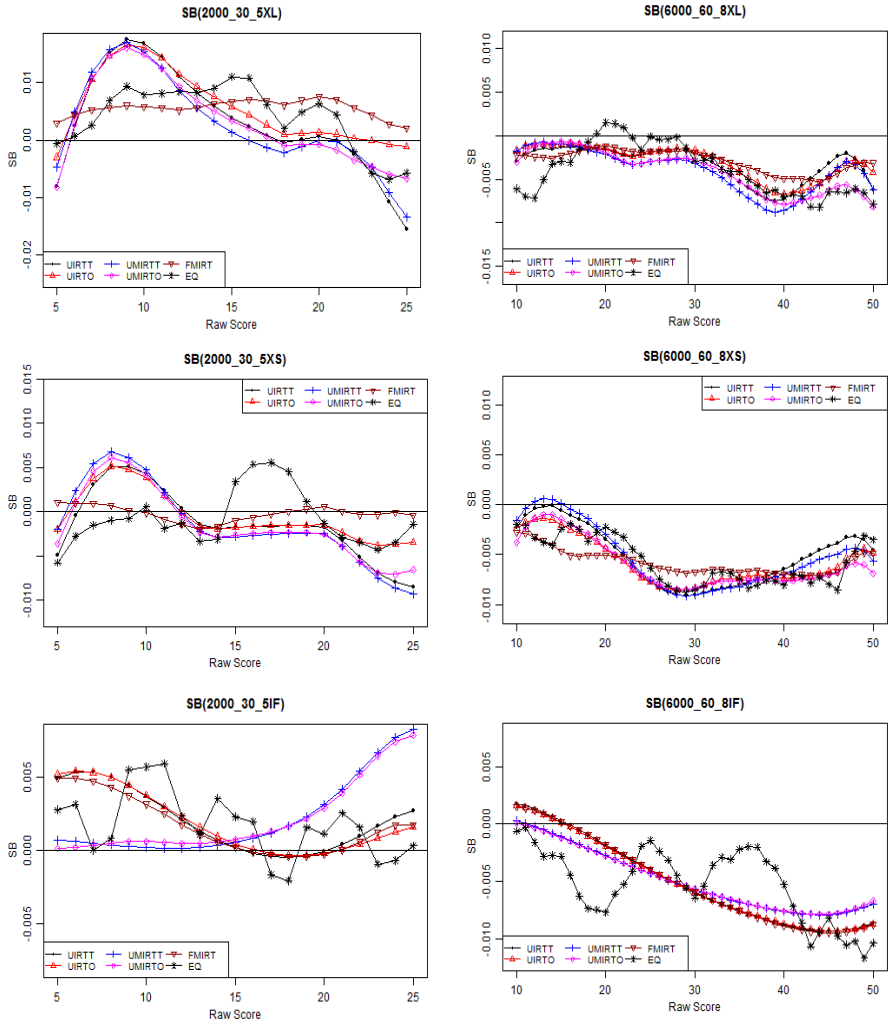

Figure 1: Standardized Bias

With respect to SEE (see Figure 2), the magnitudes of SEE for all the conditions were smaller when the sample size was large rather than small (i.e.,

6,000 rather than 2,000). SEEs for equipercentile equating were almost always the highest for all the score points across all the simulation conditions. SEEs for the two UMIRT procedures seemed to be slightly higher than for the two UIRT procedures and the FMIRT procedure at the low and high ends of score scale. In particular, this was more likely when the correlation between dimensions was low rather than high (i.e., 0.5 rather than 0.8).

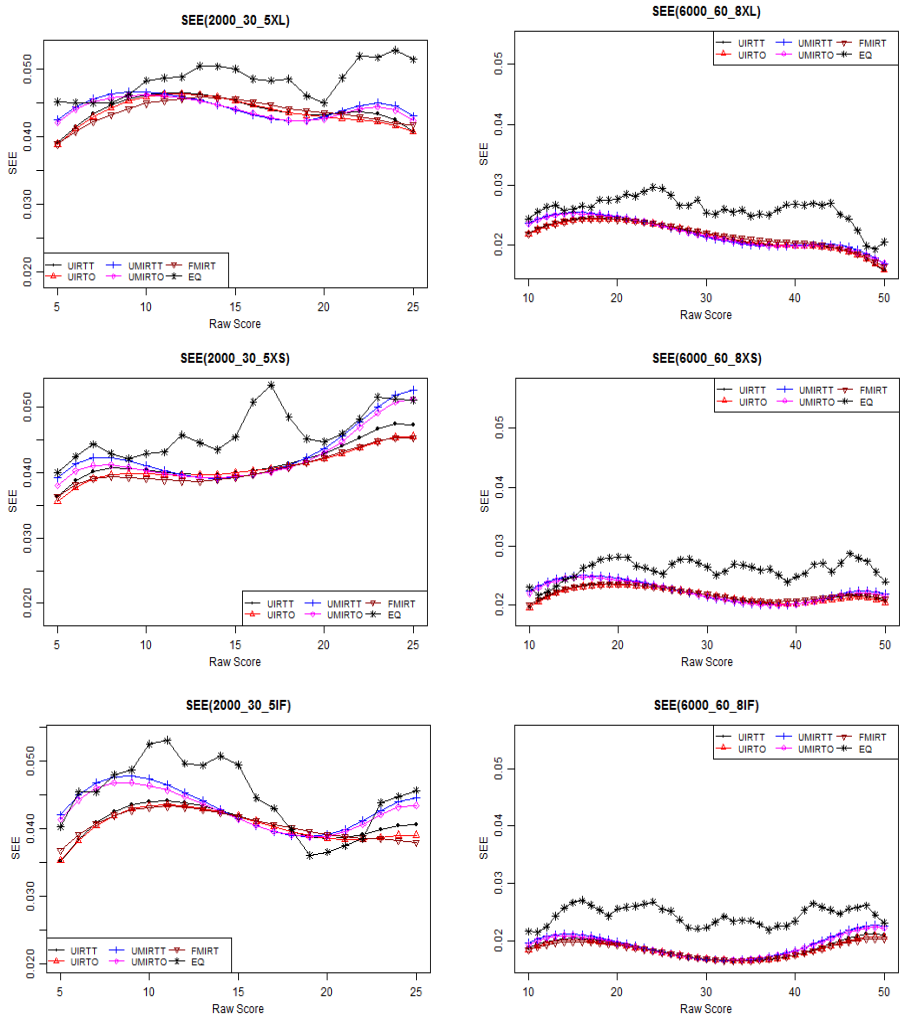

Figure 2: Standard Error of Equating

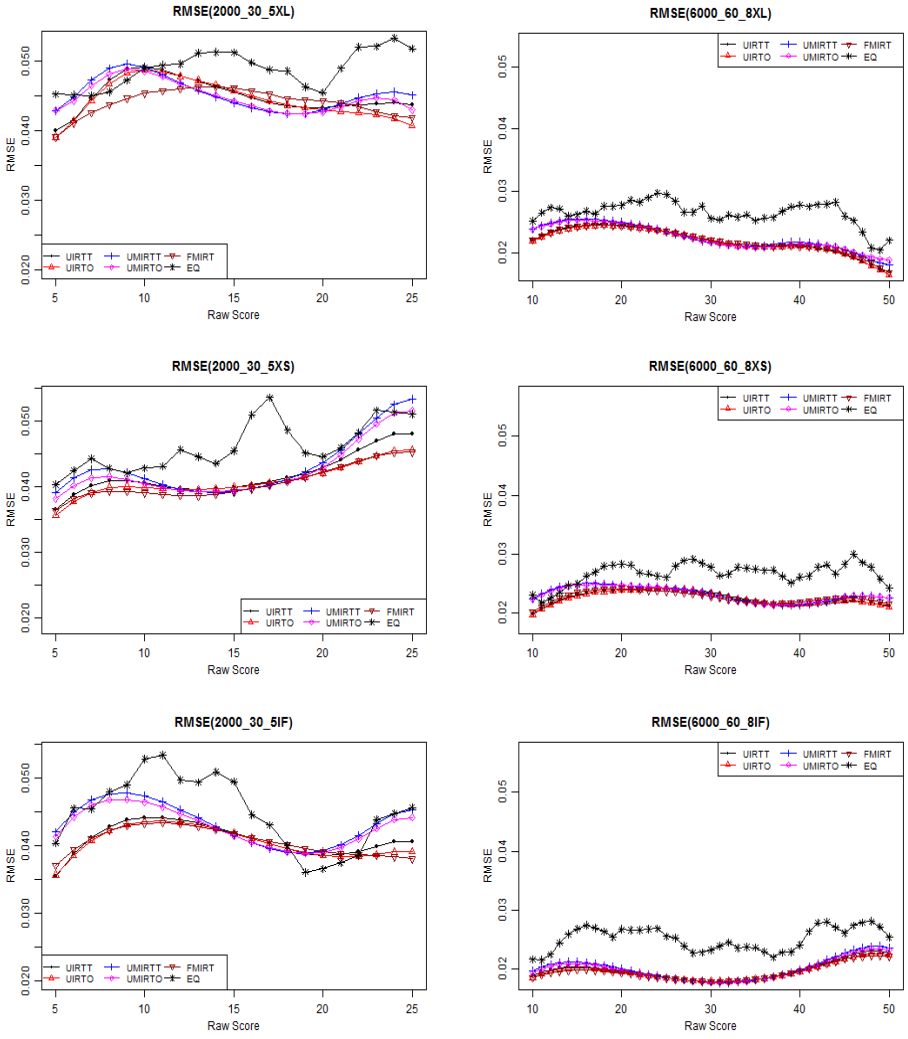

Figure 3: Root Mean Squared Error

Similar trends were observed for RMSE (see Figure 3). The magnitudes of RMSE for all the equating procedures were smaller when sample size was large rather than small (i.e., 6,000 rather than 2,000). RMSE of the EQ procedure was almost always the highest across all the simulation conditions. Additionally, in general, the two UIRT equating procedures (i.e., UIRTT and UIRTO) and the FMIRT equating procedure seemed to have lower RMSE at nearly every raw score under all the simulation conditions. When the sample size was 6,000

and the correlation between dimensions was 0.5, the lines for RMSE of FMIRT seemed to be the lowest among all the RMSE lines, especially when test length was short (i.e., the 30-item test rather than the 60-item test). However, the RMSE lines for all the equating procedures seemed to be quite overlapping throughout the middle range of the score scales.

4.2 Overall Results

For each equating procedure under each simulation condition, Table 2, Table 3, and Table 4 show the three statistics used to evaluate the overall equating results: weighted sum of absolute bias (WAB), standard error of equating (WSE), and root mean squared error (WRMSE), respectively. The first three columns in Table 2 to Table 4 indicate simulation conditions: the first column indicates test-length, the second column is for sample size, and the third column contains correlation between dimensions. The following columns indicate the values of the overall statistics for each equating procedure. Furthermore, in these Tables, UIRTT indicates UIRT true score equating, UIRTO is UIRT observed score equating, UMIRTT is unidimensionalized MIRT true score equating, UMIRTO is unidimensionalized MIRT observed score equating, FMIRT is full MIRT observed score equating and EQ is equipercentile equating.

For large form differences, the FMIRT procedure was found to perform the best with respect to WAB under five conditions: (1) sample size of 2,000, test length of 30, and correlation of 0.5; (2) sample size of 6,000, test length of 30, and correlation of 0.5; (3) sample size of 2,000, test length of 60, and correlation of 0.8; (4) sample size of 6,000, test length of 60, and correlation of 0.5; and (5) sample size of 6,000, test length of 60, and correlation of 0.8. In particular, the FMIRT procedure seemed to perform better than any other procedure when sample size was bigger (i.e., 6,000), when test was longer (i.e., 60 items), or when correlation between dimensions was lower (i.e., 0.5). Further, it is noteworthy that this FMIRT procedure was the worst equating method when large form differences were specified, when sample size was smaller (i.e., 2,000), when test was shorter (i.e., 30 items), and when correlation between dimensions was higher (i.e., 0.8).

Table 2: Weighted Sum of Absolute Bias (WAB)

N_i	N_j	ρ	UIRTT	UIRTO	UMIRTT	UMIRTO	FMIRT	EQ
XL (Large Form Differences)								
30	2,000	0.5	0.0060	0.0055	0.0053	0.0053	0.0049	0.0055
		0.8	0.0049	0.0056	0.0052	0.0053	0.0058	0.0048
	6,000	0.5	0.0078	0.0059	0.0064	0.0059	0.0031	0.0034
		0.8	0.0057	0.0048	0.0054	0.0057	0.0053	0.0061
60	2,000	0.5	0.0025	0.0020	0.0040	0.0038	0.0025	0.0040
		0.8	0.0018	0.0016	0.0026	0.0026	0.0012	0.0030
	6,000	0.5	0.0046	0.0048	0.0036	0.0036	0.0034	0.0041
		0.8	0.0026	0.0024	0.0032	0.0032	0.0022	0.0031
XS (Small Form Differences)								
30	2,000	0.5	0.0026	0.0021	0.0033	0.0030	0.0006	0.0024
		0.8	0.0067	0.0071	0.0068	0.0069	0.0072	0.0058
	6,000	0.5	0.0041	0.0033	0.0027	0.0025	0.0010	0.0017
		0.8	0.0026	0.0021	0.0024	0.0027	0.0024	0.0027
60	2,000	0.5	0.0028	0.0024	0.0030	0.0029	0.0007	0.0021
		0.8	0.0026	0.0022	0.0025	0.0020	0.0026	0.0042
	6,000	0.5	0.0031	0.0030	0.0030	0.0029	0.0012	0.0013
		0.8	0.0043	0.0048	0.0044	0.0049	0.0047	0.0047
IF (No Form Differences)								
30	2,000	0.5	0.0017	0.0016	0.0016	0.0016	0.0015	0.0020
		0.8	0.0009	0.0009	0.0016	0.0016	0.0007	0.0028
	6,000	0.5	0.0002	0.0003	0.0003	0.0003	0.0002	0.0009
		0.8	0.0005	0.0005	0.0004	0.0004	0.0005	0.0019
60	2,000	0.5	0.0027	0.0027	0.0027	0.0026	0.0030	0.0030
		0.8	0.0029	0.0029	0.0030	0.0030	0.0032	0.0040
	6,000	0.5	0.0028	0.0028	0.0028	0.0028	0.0026	0.0032
		0.8	0.0043	0.0043	0.0040	0.0040	0.0043	0.0040

For small form differences, the FMIRT procedure was also found to be the best equating procedure under four conditions: (1) sample size of 2,000, test length of 30, and correlation of 0.5; (2) sample size of 6,000, test length of 30, and correlation of 0.5; (3) sample size of 2,000, test length of 60, and correlation of 0.5; and (4) sample size of 6,000, test length of 60, and correlation of 0.5.

Specifically, the FMIRT procedure was found to perform the best when the correlation between dimensions was lower (i.e., 0.5) as all these four conditions assumed the correlation of 0.5 between dimensions. FMIRT was the worst equating method when form differences were small, sample size was small (i.e., 2,000), test length was short (i.e., 30 items), and correlation between dimensions was high (i.e., 0.8). For other conditions, meaningful patterns were not observed.

When there were no form differences, the FMIRT procedure was found to perform the best under four conditions: (1) sample size of 2,000, test length of 30, and correlation of 0.5; (2) sample size of 2,000, test length of 30, and correlation of 0.8; (3) sample size of 6,000, test length of 30, and correlation of 0.5; and (4) sample size of 6,000, test length of 60, and correlation of 0.5. In particular, the FMIRT procedure seemed to perform better than any other procedures when test was shorter (i.e., 30 items). When no form differences were assumed and test length was 30, the FMIRT procedure performed the best, except under one condition—sample size was 6,000 and correlation was 0.8 (see Table 2). Furthermore, when no form differences were assumed, the IRT-based equating procedures yielded almost the same WABs (see Table 2), and the equipercentile equating procedure yielded the highest WAB except for the one condition when sample size was 6,000, test length was 60, and correlation was 0.8.

For the WABs of the two UIRT equating procedures and the two UMIRT equating procedures (i.e., UIRTT, UIRTO, UMIRTT, and UMIRTO), when large form differences were assumed and other conditions were equal, the WAB values were larger when the correlation between dimensions was smaller (i.e., 0.5 rather than 0.8) across the simulation conditions, except for one case when sample size was 2,000 and test length was 30. By contrast, when no form differences were assumed, WABs were smaller across the simulation conditions when correlation between dimensions was smaller (i.e., 0.5), except for a sample size of 2,000 and test length of 30 (see Table 2). When small form differences were assumed, two patterns were observed: (1) WABs were smaller when the correlation was smaller under two conditions: sample size of 2,000 with test length of 30, and sample size of 6,000 with test length of 60, and (2) the WABs were larger when the correlation was smaller under two conditions: sample size of 2,000 with test length of 60, and sample size of 6,000 with test length of 30 (note, however, that UMIRTO yielded the larger WAB value when the correlation was 0.8 for a sample size of 6,000 with test length of 30; see Table 2).

Table 3 provides WSEs for each equating procedure under each simulation condition. When large form differences were specified, results showed three clear patterns. First, across all eight conditions, WSEs for equipercentile equating were the largest. This was an expected finding, since the unsmoothed equipercentile equating procedure was used in this study. Second, all five IRT-based equating procedures performed similarly with each other across all the conditions; the largest differences in the WSE values among the five IRT procedures across all the conditions ranged from .0002 to .0023 (see Table 3). Third, WSEs of all the equating procedures across all the conditions were larger when sample

size was smaller (i.e., 2,000 rather than 6,000).

Table 3: Weighted Sum of SEE (WSE)

N_i	N_j	ρ	UIRTT	UIRTO	UMIRTT	UMIRTO	FMIRT	EQ
XL (Large Form Differences)								
30	2,000	0.5	0.0411	0.0408	0.0412	0.0410	0.0408	0.0449
		0.8	0.0357	0.0354	0.0366	0.0364	0.0357	0.0404
	6,000	0.5	0.0207	0.0203	0.0225	0.0220	0.0202	0.0239
		0.8	0.0192	0.0189	0.0207	0.0204	0.0188	0.0230
60	2,000	0.5	0.0353	0.0351	0.0358	0.0356	0.0354	0.0432
		0.8	0.0389	0.0387	0.0389	0.0387	0.0387	0.0446
	6,000	0.5	0.0208	0.0207	0.0219	0.0218	0.0207	0.0249
		0.8	0.0196	0.0196	0.0199	0.0198	0.0197	0.0235
XS (Small Form Differences)								
30	2,000	0.5	0.0384	0.0378	0.0394	0.0388	0.0375	0.0427
		0.8	0.0385	0.0381	0.0399	0.0395	0.0385	0.0428
	6,000	0.5	0.0213	0.0209	0.0239	0.0234	0.0207	0.0236
		0.8	0.0201	0.0197	0.0219	0.0214	0.0197	0.0232
60	2,000	0.5	0.0378	0.0377	0.0380	0.0379	0.0374	0.0436
		0.8	0.0365	0.0364	0.0363	0.0362	0.0366	0.0431
	6,000	0.5	0.0184	0.0183	0.0186	0.0184	0.0179	0.0228
		0.8	0.0195	0.0194	0.0200	0.0199	0.0196	0.0234
IF (No Form Differences)								
30	2,000	0.5	0.0381	0.0377	0.0400	0.0395	0.0378	0.0420
		0.8	0.0361	0.0356	0.0361	0.0357	0.0355	0.0410
	6,000	0.5	0.0225	0.0223	0.0247	0.0244	0.0220	0.0252
		0.8	0.0256	0.0254	0.0268	0.0266	0.0253	0.0278
60	2,000	0.5	0.0369	0.0368	0.0375	0.0373	0.0367	0.0425
		0.8	0.0381	0.0378	0.0381	0.0378	0.0381	0.0444
	6,000	0.5	0.0229	0.0228	0.0234	0.0232	0.0224	0.0262
		0.8	0.0166	0.0165	0.0171	0.0169	0.0164	0.0219

When small form differences were specified, results also showed a similar pattern of findings: (1) equipercentile equating performed the worst, yielding the largest WSE across the conditions except when sample size was 6,000, test

length was 30, and correlation was 0.5; (2) WSEs were larger when sample size was smaller (i.e., 2,000 rather than 6,000) across all the conditions; and (3) all five IRT-based equating procedures performed similarly. It should be noted, however, that despite the performance similarity across the five IRT methods, it is notable that the FMIRT procedure was found to perform the best under five conditions: (1) sample size of 2,000, test length of 30, and correlation of 0.5; (2) sample size of 6,000, test length of 30, and correlation of 0.5; (3) sample size of 6,000, test length of 30, and correlation of 0.8; (4) sample size of 6,000, test length of 60, and correlation of 0.5; and (5) sample size of 6,000, test length of 60, and correlation of 0.8 (the WSE values ranged from .0179 to .0375 for these five conditions; see Table 3). Given that these five conditions include the four conditions for which the correlation between dimensions was set to 0.5, it can be inferred that when small form differences are assumed, the FMIRT procedure is likely to perform the best if the correlation between dimensions is low. On the other hand, when small form differences were assumed and the correlation between dimensions was larger (i.e., 0.8), the other IRT procedure (UIRTO) performed better than the FMIRT procedure (see Table 3).

A similar pattern of findings was also observed when no form differences were specified: (1) poor performance of the equipercentile equating procedure, (2) larger WSEs when the sample size was smaller, and (3) performance similarity among all five IRT-based equating procedures. Additionally, the FMIRT procedure was found to perform the best under six conditions: (1) sample size of 2,000, test length of 30, and correlation of 0.8; (2) sample size of 6,000, test length of 30, and correlation of 0.5; (3) sample size of 6,000, test length of 30, and correlation of 0.8; (4) sample size of 2,000, test length of 60, and correlation of 0.5; (5) sample size of 6,000, test length of 60, and correlation of 0.5; and (6) sample size of 6,000, test length of 60, and correlation of 0.8.

Table 4 provides WRMSEs for each equating procedure under each simulation condition. The pattern of results was similar to the result for WSEs. This is because (1) the magnitude of WRMSE is a function of both the magnitudes of WAB and WSE and (2) the magnitude of WAB was found to be relatively small in this study. Therefore, the findings regarding WRMSE can be summarized as follows. Regardless of the extent to which form differences were assumed, (1) WRMSEs of equipercentile equating were the largest, (2) all the five IRT-based equating procedures performed similarly across all the conditions, and (3) WRMSEs for all the equating procedures across all the conditions were larger when sample size was smaller (i.e., 2,000 rather than 6,000). Additionally, when small form differences or no form differences were specified, WRMSEs of the FMIRT procedure tended to be lower than those for any other methods.

Table 4: Weighted Sum of RMSE (WRMSE)

N_i	N_j	ρ	UIRTT	UIRTO	UMIRTT	UMIRTO	FMIRT	EQ
XL (Large Form Differences)								
30	2,000	0.5	0.0419	0.0416	0.0420	0.0417	0.0411	0.0454
		0.8	0.0362	0.0360	0.0371	0.0369	0.0363	0.0410
	6,000	0.5	0.0228	0.0215	0.0237	0.0231	0.0205	0.0242
		0.8	0.0205	0.0197	0.0216	0.0215	0.0197	0.0240
60	2,000	0.5	0.0355	0.0352	0.0361	0.0359	0.0355	0.0435
		0.8	0.0389	0.0387	0.0390	0.0389	0.0387	0.0449
	6,000	0.5	0.0217	0.0216	0.0224	0.0223	0.0212	0.0254
		0.8	0.0199	0.0198	0.0203	0.0202	0.0199	0.0239
XS (Small Form Differences)								
30	2,000	0.5	0.0386	0.0378	0.0396	0.0390	0.0375	0.0429
		0.8	0.0393	0.0390	0.0407	0.0403	0.0394	0.0434
	6,000	0.5	0.0219	0.0213	0.0242	0.0237	0.0207	0.0237
		0.8	0.0203	0.0198	0.0221	0.0217	0.0199	0.0235
60	2,000	0.5	0.0380	0.0378	0.0382	0.0380	0.0374	0.0437
		0.8	0.0367	0.0365	0.0364	0.0362	0.0368	0.0435
	6,000	0.5	0.0189	0.0186	0.0190	0.0187	0.0180	0.0229
		0.8	0.0203	0.0203	0.0208	0.0207	0.0203	0.0240
IF (No Form Differences)								
30	2,000	0.5	0.0381	0.0378	0.0401	0.0396	0.0379	0.0421
		0.8	0.0361	0.0356	0.0362	0.0358	0.0355	0.0412
	6,000	0.5	0.0225	0.0223	0.0247	0.0244	0.0220	0.0253
		0.8	0.0256	0.0254	0.0268	0.0266	0.0253	0.0279
60	2,000	0.5	0.0371	0.0369	0.0376	0.0374	0.0369	0.0426
		0.8	0.0382	0.0380	0.0382	0.0380	0.0383	0.0447
	6,000	0.5	0.0232	0.0230	0.0236	0.0235	0.0226	0.0265
		0.8	0.0175	0.0174	0.0178	0.0176	0.0173	0.0225

5 Discussion

The purpose of this research was to compare the performance of six equating procedures when the test is multidimensional. Using simulated data, various key

issues were addressed concerning equating mixed-format tests. In so doing, the performance of the six equating procedures was examined in terms of accuracy (i.e., bias), standard error of equating, and overall error.

Whether an equating procedure is based on a MIRT framework (i.e., UMIRTT, UMIRTO, and FMIRT) or a UIRT framework (i.e., UIRTT and UIRTO) did not necessarily yield big differences in terms of bias. The two procedures based on a UIRT framework and the two procedures based on a MIRT framework (i.e., UIRTT, UIRTO, UMIRTT, and UMIRTO) performed similarly in recovering population equating relationships established by conducting the FMIRT procedure. Furthermore, the two equating procedures based on a UIRT framework (the UIRTT and UIRTO procedures) and the one procedure based on a MIRT framework (the FMIRT procedure) performed similarly in recovering population equating relationships established by identity equating. Furthermore, with respect to WSE, all five IRT-based equating procedures performed similarly across all the conditions.

When some degree of form differences was assumed, all six equating procedures in this study performed better at recovering population equating relationships when a test becomes longer. On the other hand, when no form differences were assumed, all the equating procedures performed better for shorter tests. In particular, the FMIRT procedure seemed to perform better than any other procedures when the test was shorter and when no form differences were specified. Thus, test length influences equating performance across all the equating procedures, and when some form differences are assumed, increasing test length improved equating performance. It is unclear, however, why the opposite pattern was observed under the no-form-differences conditions.

The effect of sample size on recovering population equating relationships in terms of bias for each equating procedure seemed to be minimal. Given the fact that all the equating procedures performed quite well at recovering population equating relationships, the small sample size condition in this study (i.e., the sample size of 2,000) might not be small enough to cause an effect on equating performance. Among the six equating procedures, the equipercentile equating procedure showed the biggest improvement in accuracy (i.e., the biggest decrease in bias) when sample size became larger (although the absolute magnitude was generally small). However, with respect to random error and overall error (i.e., WSE and WRMSE), the larger sample size was associated with better performance of all the equating procedures across all the conditions. This result was expected because standard errors of equating become smaller as a sample size increases (Kolen & Brennan, 2004), and overall error in this study was governed mostly by the standard errors of equating given the small magnitude of bias.

All six equating procedures performed better at recovering population equating relationships when the form differences became smaller. The improvement in accuracy as the form differences became smaller (i.e., the differences in the magnitudes of bias for each equating procedure between the large-form-differences condition and the small-form-differences condition) were similar across all the six equating procedures. That is, as the form differences become smaller, similar degrees of improvement in recovering population equating relationships were

observed across all six equating procedures.

The effect of the correlation between dimensions on the equating results was minimal except for the case of bias yielded for the FMIRT and EQ procedures. All the other four equating procedures performed similarly under the high and low correlation conditions in terms of bias. However, when some degree of form differences was assumed, the FMIRT and EQ procedures yielded better performance in terms of bias when the correlation between dimensions was lower than higher. For the FMIRT equating results, the current finding is consistent with the Lee and Brossman's (2012) finding that the magnitude of bias in the results of simple-structure MIRT (SS-MIRT) tends to decrease as the proficiency correlation decreases. The results from the FMIRT procedure are comparable to those for the SS-MIRT procedure because both procedures are IRT observed-score equating procedures under the MIRT framework, even though the test structures for those procedures are not the same: SS-MIRT is based on the simple-structure MIRT framework, whereas FMIRT is based on the complex MIRT framework. For the EQ equating results, however, it was not expected that bias in the equating results from the EQ procedure would be higher in the high correlation condition than in the low correlation condition. One possible contributing explanation may be that the marginal distributions for Form X scores under the high correlation condition were quite different from the normal bell-shapes in most of the rest of this study. However, given that MIRT research is still in its infancy, further theoretical and empirical research is required to answer why this pattern of results was found for the EQ procedure.

The two UIRT procedures (UIRTT and UIRTO) were expected to perform better when the correlation between dimensions becomes higher as the test would be closer to unidimensional in the higher correlation condition than the lower correlation condition. This expectation was supported in terms of WSE and WRMSE, but with respect to WAB, it was supported only in the condition for the UIRTT procedure that had some degree of form differences. Little theoretical rationale is available to explain why this was supported only in this particular condition, and thus this finding requires further examination. However, the amounts of error yielded by the two UIRT procedures in the high and low correlation conditions were quite similar: the differences in WAB between these two conditions for both UIRT procedures were less than .0005, and for those in WSE and WRMSE the differences were less than .001.

Among all the equating procedures in this study, the FMIRT procedure almost always performed the best in terms of bias, standard error, and overall error. This finding was more evident especially when the correlation between dimensions was low. The FMIRT procedure is also the only procedure that uses the multidimensional item and person information in a straightforward way and, at the same time, does not assume unidimensionality. Thus, when the tests are multidimensional, using the FMIRT procedure should bring benefits in terms of yielding relatively accurate equating relationships with small equating errors without assuming unidimensionality.

Despite this merit of the FMIRT procedure, other IRT-based equating procedures might be considered as well, especially if violation of the unidimensionality

is not a big concern. This is because in this study, all the equating procedures performed quite well at recovering the population equating relationships. Note also that the IRT-based equating procedures seemed to perform better than the equipercentile equating procedure. Given these two findings, a simple IRT-based equating procedure might be preferable from practical points of view: either the UIRTT or the UIRTO procedure.

Additionally, the results in this study are less likely to support use of the UMIRT procedures: the UMIRT procedures did not seem to have a big advantage over other IRT-based equating procedures for multidimensional tests. Although the UMIRT procedures performed as well as the other IRT-based methods at recovering population equating relationships, these procedures were neither the best procedures in terms of bias, standard error, and overall error, nor the simplest procedures (they need additional computational steps to estimate unidimensional item parameters and ability distributions). Lastly, in this study, the IRT-based equating procedures seemed to perform better than the equipercentile equating procedure. It should be noted, however, that this result may be limited as only the unsmoothed equipercentile procedure was used in this study. Thus it is unclear whether the current finding would hold the same if a smoothed equipercentile procedure were used instead.

To sum up, the current study suggests that equating results are influenced by both test and examinee characteristics. It should be noted, however, that the a few findings in this study may have been impacted by a number of factors that were not examined. For example, the population equating relationships were defined differently across the no-form-differences conditions and the with-form-differences conditions. Consequently, comparisons across these two conditions may not be entirely reasonable. Additionally, only one item-parameter set was selected for all the replications for the equating methods, which may not have been optimal for all replications. Thus, future research is needed in order to strengthen the generalizability of the current findings. Such research should help develop practical guidelines that can be used in various operational testing situations.

6 References

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes.* [Computer program]. Iowa City: University of Iowa.

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

Brossman, B. G., & Lee, W. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37*, 460-481.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75*, 33-57.

Cai, L. (2012). $flexMIRT^{TM}$ version 1.88: A numerical engine for multilevel item factor analysis and test scoring. [Computer software]. Seattle, WA: Vector Psychometric Group.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed format tests using a simple structure multidimensional IRT framework. In Kolen, M. J. & Lee, W. (Ed.) *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 2).* (*CASMA Monograph Number 2.2*, pp. 118-129).

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement, 8*, 453-461.

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.

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

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

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

Thompson, T., Nering, M., & Davey, T. (1997, June). *Multidimensional IRT scale linking*. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Yon, H. (2006). *Multidimensional Item Response Theory (MIRT) approaches to vertical scaling* (Unpublished doctoral dissertation). Michigan State University, East Lansing.

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.