KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: November 26, 2015
Revision received: January 6, 2015
Accepted: March 18, 2016
OnlineFirst: April 20, 2016

www.estp.com.tr
DOI 10.12738/estp.2016.3.0218 • June 2016 • 16(3) • 715-734

Research Article

The Impact of Test Dimensionality, Common-Item Set Format, and Scale Linking Methods on Mixed-Format Test Equating*

Neşe Öztürk-Gübeş[1]
Mehmet Akif Ersoy University

Hülya Kelecioğlu[2]
Hacettepe University

Abstract

The purpose of this study was to examine the impact of dimensionality, common-item set format, and different scale linking methods on preserving equity property with mixed-format test equating. Item response theory (IRT) true-score equating (TSE) and IRT observed-score equating (OSE) methods were used under common-item nonequivalent groups design. The equity property was evaluated based on first-order equity (FOE) and second-order equity (SOE) properties. A simulation study was conducted based on actual item parameter estimates obtained from the TIMSS 2011 8th grade mathematics assessment. The results showed that: (i) The FOE and SOE properties were best preserved under the unidimensional condition, were poorly preserved when the degree of multidimensionality was severe. (ii) The TSE and OSE results, which were provided by using a mixed-format common-item set, preserved FOE better compared to equating results, which provided only a multiple-choice common item set. (iii) Under the unidimensional and multidimensional test structure, characteristic curve methods performed significantly better than moment scale linking methods in terms of preserving FOE and SOE properties.

Keywords

Test equating • Mixed-format tests • First-order equity • Second-order equity • Item response theory

* This article is an extension of corresponding author's doctoral thesis at Hacettepe University. The authors would like to thank Dr. Benjamin James Andrews and Dr. Insu Paek for their invaluable contribution, insight, and support toward this research.

1 **Correspondence to:** Neşe Öztürk-Gübeş (PhD), Department of Educational Sciences, Faculty of Education, Mehmet Akif Ersoy University, Burdur Turkey. Email: nozturk@mehmetakif.edu.tr

2 Department of Educational Sciences, Hacettepe University, Ankara Turkey. Email: hulyaebb@hacettepe.edu.tr

Citation: Öztürk-Gübeş, N., & Kelecioğlu, H. (2016). The impact of test dimensionality, common-item set format, and scale linking methods on mixed-format test equating. *Educational Sciences: Theory & Practice, 16,* 715-734.

Many testing programs use mixed-format tests, which consist of multiple-choice (MC) and constructed-response (CR) items. For example, international testing programs, such as Trends International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), Programme for International Student Assessment (PISA) and national testing programs such as College Board's Advanced Placement (AP) examinations, National Assessment of Educational Progress (NEAP), Florida Comprehensive Assessment Test (FCAT) and so on. While MC items require examinees choose the response from a list of options, CR items require examinees to generate their own response (Martinez, 1999). Both item formats have their own strengths and weaknesses. A broad range of content can be tested with MC items. In addition, the responses can be scored by machine effectively and objectively (Livingston, 2009). Although MC items can be constructed to measure higher-order thinking skills, a full range of higher-order thinking processes cannot be measured with MC items. This skill, however, is represented in CR items (Balch, 1964; Messick, 1993). On the other hand, CR items have been criticized for covering a narrow range of content, being more time-consuming, expensive to score, and subjectively scored (Livingston, 2009). The rationale for using mixed-format tests is to take advantage of and eliminate the disadvantages of both item formats.

In many testing programs, there is no single form or version of the test (Braun & Holland, 1982). Because of security problems or test administration being done at different times and/or different locations, more than one test form of a test is required (Kolen & Brennan, 2004). Although test developers try to construct test forms as similarly as possible to one another, they will still differ in their difficulty (Kolen, 1984). To use scores from different forms of a test interchangeably, test scores should be equated. The purpose of equating is to establish an effective equivalence scores on two test forms such that scores from each test can be used as if they come from the same test (Petersen, 2007).

Equating is an empirical procedure (Dorans, Moses, & Eignor, 2011), because it requires a design for data collection and a rule for transforming scores on one test form to scores on another. Three data collection designs are commonly used for equating: single group design, random groups design and common-item nonequivalent groups (CINEG) design. The focus of current study is on the CINEG design. For the CINEG design, the groups of examinees which taking different test forms of a test are not assumed to be equivalent in proficiency. To disentangle the group difference from the form difference a common-item set is used for equating two forms (Kolen & Brennan, 2004).

IRT Test Equating

IRT methods are an important component of equating methodology. Many testing programs use IRT equating methods (Kolen & Brennan, 2004). IRT equating generally involves three steps: item calibration, scale transformation, and equating.

In the first step, item parameters on the different forms are estimated via concurrent calibration, or separate calibration. If the equating is conducted under CINEG design, the parameters from different forms need to be on the same IRT scale. In a concurrent calibration, item parameters on both test forms are estimated jointly in one computer run and the estimated parameters are automatically on the same scale. In a separate calibration, item parameters for each form are estimated in a single computer run. When a separate calibration is conducted, the estimated parameters are on different scales and a scale transformation or linking is needed. The purpose of scale transformation is to find two linking coefficients, such as *A* for slope and *B* for intercept. If we define Scale I and Scale J as three-parameter logistic IRT scales that differ by a linear transformation, the Θ values and item parameters for two scales are related as follows (Kolen & Brennan, 2004):

$$\theta_{Ji} = A\theta_{Ii} + B$$

$$a_{Jj} = a_{Ij} / A$$

$$b_{Jj} = Ab_{Ij} + B$$

$$c_{Jj} = c_{Ij}$$

Where *A* and *B* are constants, and are values of Θ for individual I on Scale J and Scale I. The item parameters for item j on Scale J are , , and ; the item parameters for item j on Scale I are , , and .

Kim and Lee (2006) indicated that there are four scale linking methods which are applicable to mixed-format tests: the mean/sigma (Marco, 1977), mean/mean (Loyd & Hoover, 1980), Haebara (1980), and Stocking and Lord (1983). Mean/mean and mean/sigma methods are referred to as moment methods while the Haebara and, Stocking and Lord methods are referred to as characteristic curve methods (Kolen & Brennan, 2004). The mean/sigma method uses the means and standard deviations of item difficulty estimates from the common items to determine linking coefficients. The mean/mean method uses mean of slope and difficulty parameter estimates from the common items to determine *A* and *B* linking coefficients. Another approach for calculating the linking coefficients are characteristic curve methods. These methods are based on minimizing a loss function that depends on the metric of test calibration (Baker & Al-Karni, 1991). Characteristic curve methods consider all parameters simultaneously to find linking coefficients that minimize differences in the characteristic curve between tests. The Haebara method uses the difference between the item characteristic curves and takes the sum of the squared difference between the item characteristic curves for each item for examinees of a particular ability. However, in Stocking-Lord, the summation is taken over items for each set of parameter estimates before squaring (Kolen & Brennan, 2004).

After estimating parameters and placing them on the same scale, the next step is IRT equating. In IRT true-score equating, the true score on one form associated with a given Θ is considered to be equivalent to the true score on another form associated with that Θ. IRT true-score equating use item parameter estimates to produce estimated true score relationship. Then the estimated true score conversion is applied to the observed scores. A second IRT equating method is IRT observed-score equating. In IRT observed-score equating, IRT models are used to produce an estimated distribution of observed-number correct scores on each form, and then the observed scores are equated using an equipercentile equating method (Kolen & Brennan, 2004).

Evaluating Equating Results

After equating, the last, and one of the most important questions, is: Has it been done well enough? To answer this question, equating results should be evaluated. Harris and Crouse (1993) provided an extensive review of equating criteria used in equating studies. Lord's (1980) equity property is one of criteria which have been used to evaluate equating results. According to Lord's equity property (Kolen & Brennan, 2004, p. 10):

…It must be a matter of indifference to each examinee whether Form X or Form Y is administered…

In other words, if for examinees with a given true score, the distribution of equated scores on the new form is identical to the distribution of the scores on the old form, then Lord's equity property holds. This is not possible unless two forms are strictly parallel, in which case equating is unnecessary (Brennan, 2010; Harris, 1993). For this reason, Morris (1982) suggested a less strict definition of equity, which is called weak equity or first-order equity. According to the first-order equity (FOE) property (Morris, 1982, p. 171):

…Each individual in the test population has the same expected score on both tests…

The FOE property says that the examinees with a given ability have the same mean of equated scores on the new form as they have on the reference form or old form. FOE property can be evaluated by calculating a D_1 index, which is proposed by Tong and Kolen (2005):

$$D_1 = \frac{\sqrt{\sum_i q_i \{E[Y|\theta_i] - E[eq_Y(x)|\theta_i]\}^2}}{SD_Y}$$

where $E[Y|\theta_i]$ is the old form conditional mean for a given proficiency θ_i, $E[eq_Y(x) \mid \theta_i]$ is the conditional mean of equated score for a given proficiency θ_i, q_i is the quadrature weight at θ_i, and SD_Y is the standard deviation of Form Y. In this study, 40 quadrature points, with weights from a univariate normal distribution, were used.

Morris (1982) also suggested second-order equity (SOE) property, which requires that a standard error of measurement (SEM) conditional on true score is the same across forms after equating. SOE property can be evaluated by using the index D_2 (Tong & Kolen, 2005):

$$D_2 = \frac{\sqrt{\sum_i q_i (SEM_Y|\theta_i - SEM_{eq_Y(x)}|\theta_i)^2}}{SD_Y}$$

where $SEM_Y \mid \theta_i$ denotes conditional SEM for the old form for the examinees with proficiency θ_i, and $SEM_{eqY}(x) \mid \theta_i$ denotes conditional SEM for the equated new form for the examinees with proficiency θ_i.

Evaluating equating results with using equity criteria is important because it is directly related with special case of Lord's equity definition of equating (Harris, 1993). Equity criteria allows to assess the degree to which a given examinee is advantaged or disadvantaged by being administered alternative test forms (Bolt, 1999). While FOE property focuses on test fairness, the SOE property focuses on measurement precision (He, 2011). If FOE and SOE properties are not preserved, interchangeability of test scores after equating could not maintained.

The Purpose and Significance of the Study

Mixed-format tests also need to be equated if the aim is to use scores from different test forms interchangeably. But the use of different item formats in one test can bring some challenges to the equating process. MC and CR subtests of the same content may measure different latent characteristics, and that can cause multidimensionality due to format effects (Traub, 1993). It is known that multidimensionality is one of factors that affecting IRT equating. Since multidimensional IRT test equating is more complex, many testing programs employ unidimensional IRT equating regardless of underlying test structure (Cao, 2008). It is important to investigate robustness of unidimensional IRT equating methods when test structure is multidimensional due to format effects.

The other issue in mixed-format test equating while equating under CINEG design is composition of a common-item set. The choice of common-item set is very important in terms of quality for equating tests (Sinharay & Holland, 2007). As Kolen and Brennan (2004, p. 19) indicated:

…common-item set should be a "mini" version of the total test form…

However, in practice, although the total test includes both item formats (MC and CR items), because of some reasons (reliability, rater effect, easier to memorize CR items than MC items etc.) usually a common-item set is comprised of only MC items

(He, 2011). Should we include CR items into the common-item set? The answer of this question is ambigious in the literature. Therefore, it is valuable to examine common-item set format effect based on equity property in mixed-format test equating.

It is known that characteristic curve scale linking methods produce more accurate results than the moment methods (Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Kim & Cohen, 1992; Kim & Kolen, 2006; Kim & Lee, 2004; Ogasawara, 2001). Could we generalize this evident to mixed-format test equating? No research has investigated the performance of traditional linking methods based on equity property in mixed-format test equating. One of the purposes of this study is to investigate the relative performance of traditional linking methods in mixed-format test equating based on equity property. More specifically, the purpose of this study is to investigate the effects of dimensionality, common-item set format, and scale linking methods on mixed format test equating.

Method

Data and Construction of Mixed-Format Tests

A simulation study was conducted. To mimic a real data situation, examinee responses were simulated based on actual parameter estimates obtained from the TIMSS 2011 8th grade mathematics assessment, which are available from the TIMSS 2011 Technical Report (Martin & Mullis, 2012). For this study, a total of 50 mathematics items including 40 MC items and 10 CR items were selected from the 194 items. In TIMSS, there are two types of constructed-response items (Mullis & Martin, 2011): 1-point constructed-response items which scored as correct (1 score point) or incorrect (0 score points); 2-point constructed-response items which scored as fully correct (2 score points), partially correct (1 score point), or incorrect (0 score points). In this study, we used parameter estimates of 2-point constructed response items. Two test forms (X and Y) for equating were considered. The old form was referred as base form while the new form was referred as the target form (Kim, 2004). In this study, the old form was Form Y and new form was Form X. Each test form consisted of a unique and a common-item set. At the 8th grade, each booklet had 30 mathematics items, and at least half of the total number of points represented by all the questions came from MC items. Therefore, each test form was constructed with 24 MC and 6 CR items.

Special care was taken that the formation of alternate forms were similar as possible in terms of content and statistical characteristics. As indicated in the TIMSS 2011 report, at the eighth grade, the content domains and their percentages are: number (30%), algebra (30%), geometry (20%), and data-chance (20%). Table 1 shows the distribution of number of items for each content domain and item formats.

Table 1

The Distribution of Number of Items for Each Content Domain and Item Format

Content Domain	Form X	Form Y	Common Item Set
Number	6 (5MC+1CR)	6 (5MC+1CR)	3 (2MC+1CR)
Algebra	6 (5MC+1CR)	6 (5MC+1CR)	3 (2MC+1CR)
Geometry	4 (3MC+1CR)	4 (3MC+1CR)	2 (MC)
Data and Chance	4 (3MC+1CR)	4(3MC+1CR)	(2MC)
Total	20 (16MC+4CR)	20 (16MC+4CR)	10 (8MC+2CR)

Note. MC = multiple-choice items; CR = constructed-response items

In operational testing, although the usual target was to make the new form of the same difficulty as the old form, because of unforeseen reasons, the new form is generally easier or more difficult than the old form. In this study, we increased the mean of difficulty parameter by the amount of .22 (Sinharay & Holland, 2007) and the mean difficulty for the old and new forms are respectively .50 and .72, so the new form is more difficult than the old form. Units were in standard deviation of ability.

Factors Investigated

The three factors were considered for the simulation study and the combination of these three factors led to 24 simulation conditions [3 (levels of multidimensionality) x 2 (types of common-item set) x 4 (types of linking). In each condition, 100 replications were used.

Four levels of multidimensionality. In this study, the multidimensional test structure was constructed based on format effects. We assumed that MC items were measuring ability θ_1 and CR items were measuring ability θ_2. To specify multidimensionality based on format effect, we supposed bivariate normal (BN) distribution for latent variables. Mathematically, we could define it as $(\theta_1, \theta_2) \sim BN(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, where μ_1 and σ_1 were mean and standard deviation of σ_1, μ_2 and were those of θ_2, and ρ was the correlation coefficient between θ_1 and θ_2. To compare results of this study with other studies (Andrews, 2011; Kim & Kolen, 2006) we set correlation between θ_1 and θ_2 as .50, .80, and 1.00. In this study, the correlation value of 1.00 represented unidimensional case and there was no format effect, and the correlation value of .50 represented that there were severe format effects.

Two types of common-item format. Under this factor, we built two conditions: format representativeness and format non-representativeness. In the format representativeness case, the ratio of MC items to CR items 4:1 was reflected to the common-item set, which resulted in 8 MC and 2 CR items. In the format non-representativeness case, the common-item set consisted of only 10 MC items (Wolf, 2013).

Four types of linking methods. Under this factor, we considered two moment linking methods (mean/sigma and mean/mean) and two characteristic curve linking methods (Stocking-Lord and Haebara) which were extended to mixed format tests by Kim and Lee (2006).

Data Generation and Procedures

This study was conducted under CINEG design. Two groups of examinees were considered: Group 1 and Group 2. Examinees in Group 1 were associated with Form X (new form) and examinees in Group 2 were associated with Form Y (old form). It was assumed that the ability distribution of two groups was not equivalent and examinees in Group 2 were more competent than those in Group 1. The item responses for Group 1 were generated from , and for Group 2 they were generated from . For each group, 3000 examinees' responses were generated.

TIMSS uses IRT scaling methods to describe students' achievement and provide accurate measures of trends. For reflecting TIMSS in this study, item responses for two forms were generated based on the three-parameter logistic model (3PL; Birnbaum, 1968) model for MC items and generalized partial credit model (GPC; Muraki, 1992) for CR items. Data generation was conducted by using the SimulateRwo. bat in the SimuMIRT (Yao, 2003) program. MC item responses were simulated by using a multidimensional three-parameter logistic model (M-3PL; Yao & Schawarz, 2006). For a dichotomous item, j, the probability of a correct response to item j for an examinee with ability for the M-3PL model is

$$P_{ij1} = P\left(x_{ij} = 1 \middle| \vec{\theta_i}, \vec{\beta_j}\right) = \beta_{3j} + \frac{1-\beta_{3j}}{1+e^{(-\vec{\beta}_{2j} \odot \vec{\theta}_i^T + \beta_{1j})}},$$

where

x_{ij} = 0 or 1 is the response of examinee *i* to item *j*.

$\vec{\beta}_{2j} = (\vec{\beta}_{2j1}, \ldots, \vec{\beta}_{2jD})$ is a vector of dimension D for item discrimination parameters.

β_{1j} is the scale difficulty parameter.

β_{3j} is the scale guessing parameter.

$\vec{\beta}_{2j} \odot \vec{\theta}_i^T$ is a dot product of two vectors.

CR item responses were generated with regard to the multidimensional version of the generalized partial credit model (M-2PPC, Yao & Schawarz, 2006). For a polytomous item, j, the probability of a response k-1 to item j for an examinee with ability $\vec{\theta}_i^T$ for the M-2PPC is

$$P_{ijk} = P(x_{ij} = k-1 | \vec{\theta_i}, \vec{\beta_j} = \frac{e^{(k-1)\vec{\beta}_{2j} \odot \vec{\theta}_i^T - \sum_{t=1}^{k} \beta_{\delta_t} j}}{\sum_{m=1}^{K_j} e^{((m-1)\vec{\beta}_{2j} \odot \vec{\theta}_i^T - \sum_{t=1}^{m} \beta_{\delta_t} j}},$$

Where

$X_{ij} = 0, \ldots, K_{j-1}$ is the response of examinee I to item j.

$\vec{\beta}_{2j} = (\vec{\beta}_{2j1}, \ldots, \vec{\beta}_{2jD})$ is a vector of dimension D for item discrimination parameters.

$\beta_{\delta_k j}$ for $k = 1,2, \ldots, K_j$ are thereshold parameters, $\beta_{\delta_1 j} = 0$, and K_j is the number of response categories for the th item.

After generating data, unidimensional IRT calibration was done for Form X and Form Y separately, assuming a 3PL model for the MC items and GPC model for CR items. The computer program PARSCALE (Muraki & Bock, 2003) was used for IRT calibration. After estimating item and ability parameters, four linking methods were used to transform the estimated item parameters on the new form scale to old form scale. The computer program STUIRT (Kim & Kolen, 2004) was used for scale linking. After placing item and ability parameters on a common scale, IRT true-score and observed-score equating was conducted using POLYEQUATE (Kolen, 2004a).

Equating results were evaluated based on FOE and SOE properties. Both equity properties are conditional on a proficiency level. To evaluate equating results based on equity properties, a psychometric model or models are assumed (Tong & Kolen, 2005). In this study, a 3PL model was assumed for MC items and a GPC model was assumed for CR items. For evaluating FOE and SOE property D_1 and D_2 indexes were calculated. The conditional expected scale scores and conditional SEMs which were required for the calculating D_1 and D_2 indexes were obtained from POLYCSEM (Kolen, 2004b) computer program. To analyze all 100 datasets generated for each condition, these four computer programs were operated using R software with batch files. Then simulation factors were evaluated based on the mean of the D_1 and D_2 values over all 100 replications. A large D_1 and D_2 value suggested that FOE and SOE properties are not preserved sufficiently (Tong & Kolen, 2005). In addition, an analysis of variance (ANOVA) was performed to determine factors that had significant effects.

Results

The effects of factors on TSE and OSE results were evaluated based on mean of D_1 ($\overline{D}_1$) and D_2 ($\overline{D}_2$) values over 100 replications for each condition. Table 2 and Table 3 showed the $\overline{D}_1$ and $\overline{D}_2$ values. In order to examine the factors and see if their interaction had a statistically meaningful effect on IRT TSE and OSE results, a 3-way ANOVA was done. Cohen's (1988) guidelines of effect size in ANOVA (partial η^2 = small: .01, medium: .06, large: .14) were used. Results from the 3-way ANOVA were presented in Table 4. The two-way interactions which were statistically significant effects on equating results were presented in Figure 1 and Figure 2.

Table 2

Mean of D_1 Values for the IRT True-Score Equating and Observed-Score Equating

		Scale Linking Methods							
		M-M		M-S		SL		HA	
Anchor	Correlation	TSE	OSE	TSE	OSE	TSE	OSE	TSE	OSE
	1.00	.115	.112	.103	.100	.076	.072	.087	.082
FR	.80	.177	.172	.142	.130	.105	.088	.125	.110
	.50	.462	.541	.335	.308	.250	.160	.307	.265
	1.00	.125	.124	.128	.139	.076	.072	.082	.078
FNR	.80	.227	.228	.213	.215	.134	.122	.141	.129
	.50	.617	.723	.578	.663	.338	.305	.347	.321

Note. FR= format representativeness; FNR= format non-representativeness; M-M= Mean-Mean; M-S: Mean-Sigma, SL= Stocking-Lord, HA= Haebara

Table 3

Mean of D_2 Values for the IRT True-Score and Observed-Score Equating

		Scale Linking Methods							
		M-M		M-S		SL		HA	
Anchor	Correlation	TSE	OSE	TSE	OSE	TSE	OSE	TSE	OSE
	1.00	.038	.031	.038	.027	.044	.030	.042	.031
FR	.80	.046	.039	.038	.027	.048	.032	.044	.032
	.50	.176	.142	.077	.047	.097	.057	.085	.053
	1.00	.037	.030	.059	.036	.044	.030	.043	.030
FNR	.80	.040	.034	.053	.037	.041	.028	.040	.028
	.50	.128	.100	.111	.083	.078	.048	.076	.047

Note. FR= format representativeness; FNR= format non-representativeness; M-M= Mean-Mean; M-S: Mean-Sigma, SL= Stocking-Lord, HA= Haebara

Table 4

ANOVA Results for First-Order Equity and Second-Order Equity Properties

		TSE D_1		TSE D_2		OSE D_1		OSE D_2	
Effect	*df*	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
Multidimensionality (M)	2	1666.93*	0.41	813.30*	0.25	1334.07*	0.36	638.21*	0.21
Common-Item Format (F)	1	185.36*	0.04	0.27	0.00	265.14*	0.05	0.56	0.00
Scale Linking Methods (SLM)	3	191.40*	0.11	48.25*	0.03	289.89*	0.15	146.76*	0.08
M*F	2	48.22*	0.02	17.41*	0.01	75.61*	0.03	9.77*	0.004
M*SLM	6	43.79*	0.05	64.25*	0.08	91.50*	0.10	108.79*	0.12
F*SLM	3	31.28*	0.02	85.77*	0.05	44.39*	0.03	78.14*	0.05
M*F*SLM	6	4.20*	0.01	9.12*	0.01	6.86*	0.01	17.30*	0.02
Error	4776								
Total	4799								

Note. $*p < .05$, TSE = true-score equating, OSE = observed-scored equating.

Multidimensionality

In Table 2, under both of the equating methods, the smallest values were provided when tests were unidimensional and the largest value was provided when multidimensionality was severe (in other words, when the correlation between abilities was .50). As the correlation between abilities decreased, values increased and preserving FOE equity property decreased. Also, the 3-way ANOVA results in Table 4 showed that multidimensionality had a significant and large effect on IRT TSE and OSE equating results in terms of preserving FOE property [$F_{TSE}(2,4776) = 1666.93$, $p < .05$, $\eta^2 = .41$; $F_{OSE}(2,4776) = 1334.07$, $p < .05$, $\eta^2 = .36$]. All pairwise comparisons based on the D_1 index were statistically significant and as the correlation between abilities decreased, the marginal means increased (see Table 5).

Table 5

Pairwise Comparisons for Multidimensionality

	TSE D_1		TSE D_2		OSE D_1		OSE D_2	
Correlations	M	SE	M	SE	M	SE	M	SE
1.00 -0.80	-.040*	.002	.000	.001	-.035*	.002	.000	.000
1.00-0.50	-.186*	.002	-.031*	.001	-.185*	.002	-.021*	.000
0.80-0.50	-.146*	.002	.031*	.001	-.150*	.002	-.020*	.000

Note. *$p < .05$, TSE = true score equating, OSE = observed scored equating, M = mean difference, SE = standart error.

As seen in Table 3, TSE and OSE results had the smallest values under the unidimensional condition or when multidimensionality was not severe (in other words, when the correlation between abilities was .80). ANOVA results showed that multidimensionality had a significant and large effect on TSE and OSE results in terms of preserving SOE property [$F_{TSE}(2,4776) = 813.30$, $p < .05$, $\eta^2 = .25$; $F_{OSE}(2,4776) = 638.21$, $p < .05$, $\eta^2 = .21$]. Comparisons among degree of multidimensionality showed that based on the SOE property for both equating results, except for 1.00–.80 comparisons, all pairwise comparisons were significant and as the degree of multidimensionality increased mean of D_2values increased (See Table 5). We can say that TSE and OSE results best preserved SOE property when tests were unidimensional or multidimensionality was not severe.

Common-Item Format

Table 2 showed that compared to format non-representativeness condition, except for unidimensional/HA case, TSE and OSE results had lower or equal values when the common-item set was format representativeness. The 3-way ANOVA results showed that the common-item format had a significant and small effect on TSE and OSE results [$F_{TSE}(1,4776) = 185.36$, $p < .05$, $\eta^2 = .04$; $F_{OSE}(1,4776) = 265.14$, $p < .05$, $\eta^2 = .05$]. Based on pairwise comparisons, the mean differences between values, which provided from representativeness and non-representativeness conditions, were statistically significant ($\Delta\mu_{TSE}$=-.038, $\Delta\mu_{OSE}$ = -.051, p < .05). Both TSE and OSE results had lower $\overline{D}_1$ values when a common-item set represented the total test item format.

As you see in Table 3, TSE and OSE results, which were provided from common-item format representativeness and non-representativeness conditions, generally (except for M-S condition) format non-representative (FNR) cases had smaller $\overline{D}_2$ values compared to format representative (FR) cases. In contrast, equating results under M-S scale linking method had the smallest $\overline{D}_2$ values when the common-item set represented the total test item format. But the ANOVA results showed that the common-item format had no significant effect on both TSE and OSE results based on SOE criteria [$F_{TSE}(1,4776) = .273$, $p > .05$, $\eta^2 = .00$; $F_{OSE}(1,4776) = .562$, $p > .05$, $\eta^2 = .00$]. We can say that including or excluding CR items to common-item set did not make any statistical difference based on SOE property.

Scale Linking Methods

Regardless of multidimensionality and common-item format, TSE and OSE results had lower $\overline{D}_1$ values when the characteristic curve linking methods were used (See Table 2). The 3-way ANOVA results showed that scale linking methods had a medium effect on TSE and a large effect on OSE results in terms of preserving FOE property [$F_{TSE}(3,4776) = 191.40$, $p < .05$, $\eta^2 = .11$; $F_{OSE}(3,4776) = 289.89$, $p < .05$, $\eta^2 = .15$]. The pairwise comparisons (see Table 6) indicated that the largest mean difference was between MM and SL methods under both equating methods. As you see in Table 6, the characteristic curve methods had lower values compared to moment methods for TSE and OSE results.

Table 6

Pairwise Comparisons of Scale Linking Methods

	TSE D_1		TSE D_2		OSE D_1		OSE D_2	
Methods	M	SE	M	SE	M	SE	M	SE
MM vs MS	.015*	.004	.004*	.001	.025*	.004	.009*	.001
MM vs SL	.078*	.004	.009*	.001	.109*	.004	.014*	.001
MM vs HA	.067*	.004	.011*	.001	.093*	.004	.014*	.001
MS vs SL	.063*	.004	.005*	.001	.085*	.004	.005*	.001
MS vs HA	.052*	.004	.007*	.001	.069*	.004	.005*	.001
SL vs HA	-.011*	.004	.002*	.001	-.016*	.004	.000	.001

Note. *$p < .05$, TSE = true score equating, OSE = observed score equating, M = mean difference, SE = standart error, MM = mean/mean, MS = mean/sigma, SL = Stocking-Lord, HA = Haebara.

The 3-way ANOVA results showed that scale linking methods had a small effect on TSE results and medium effect on OSE results based on SOE criteria [$F_{TSE}(3,4776) = 48.25$, $p < .05$, $\eta^2 = .03$; $F_{OSE}(3,4776) = 146.76$, $p < .05$, $\eta^2 = .08$]. As seen in the pairwise comparisons in Table 6 while the TSE results had the largest mean difference between MM and HA methods, the OSE results had the largest mean difference between MM-SL or MM-HA methods. Pairwise comparisons in Table 6 also indicated that HA and SL methods had the same mean and there was no significant difference between these two characteristic curve methods. For both equating methods, we can say that characteristic

curve scale linking methods had smaller values compared to moment methods and SOE property was preserved well with characteristic curve linking methods.

Interaction of Factors

*Multidimensionality*common-item format interaction* had statistically significant and a small affect on TSE and OSE results based on FOE criteria [$F_{TSE}(2,4776) = 48.22$, $p < .05$, $\eta^2 = .02$; $F_{OSE}(2,4776) = 75.61$, $p < .05$, $\eta^2 = .03$]. As seen in Figures 1a and 1d, when data were unidimensional two common-item format yielded similar D_1 values, but as the correlation between abilities was decreased, the FR common-item condition always had lower D_1 values than FNR conditions. Based on SOE property, while the *multidimensionality*common-item format* interaction had a statistically significant and small affect on TSE results, it had statistically significant but practically not meaningful effect on OSE results [$F_{TSE}(2,4776) = 17.41$, $p < .05$, $\eta^2 = .01$; $F_{OSE}(2,4776) = 9.77$, $p < .05$, $\eta^2 = .004$]. Figures 2a and 2d showed that under unidimensional test structure, TSE and OSE results had lower D_2 values when common-item set was represented the total test item format. However, as the test structure became multidimensional both equating results tended to have lower D_2 values when common-item set was comprised of only multiple-choice items.

*Multidimensionality*scale linking methods interaction* had a statistically significant and small effect on TSE results and a medium effect on OSE results in terms of preserving FOE property [$F_{TSE}(6,4776) = 43.79$, $p < .05$, $\eta^2 = .05$; $F_{OSE}(6,4776) = 91.50$, $p < .05$, $\eta^2 = .10$]. As seen in Figures 1b and 1e, under unidimensional and multidimensional data structure characteristic curve methods had smaller D_1 values compared to moment methods and the two characteristic curve scale transforming methods Haebera and SL performed smilar in terms of preserving FOE. Further, the MM method had the largest D_1 value among the linking methods.

The *multidimensionality*scale linking methods interaction* also had a statistically significant and medium effect on both equating methods results based on SOE property [$F_{TSE}(6,4776) = 64.25$, $p < .05$, $\eta^2 = .08$; $F_{OSE}(6,4776) = 108.79$, $p < .05$, $\eta^2 = .12$]. Figures 2b and 2e showed that either the data were unidimensional or the degree of multidimensionality was not severe (.80), with one exception, the three scale linking methods (MM, SL, and HA) had similar D_2 values. The exception occurred with MS method, it had the largest D_2 values compared to other methods under unidimensional data structure. When the correlation between abilities was .50 or multidimensionality was severe, characteristic curve methods had consistently lower D_2 values compared to moment methods.

*Common-item format*scale linking methods interaction* had statistically significant and small effect on both TSE and OSE results based on FOE property [$F_{TSE}(3,4776) = 31.28$, $p < .05$, $\eta^2 = .02$; $F_{OSE}(3,4776) = 44.39$, $p < .05$, $\eta^2 = .03$]. Figure 1c and Figure

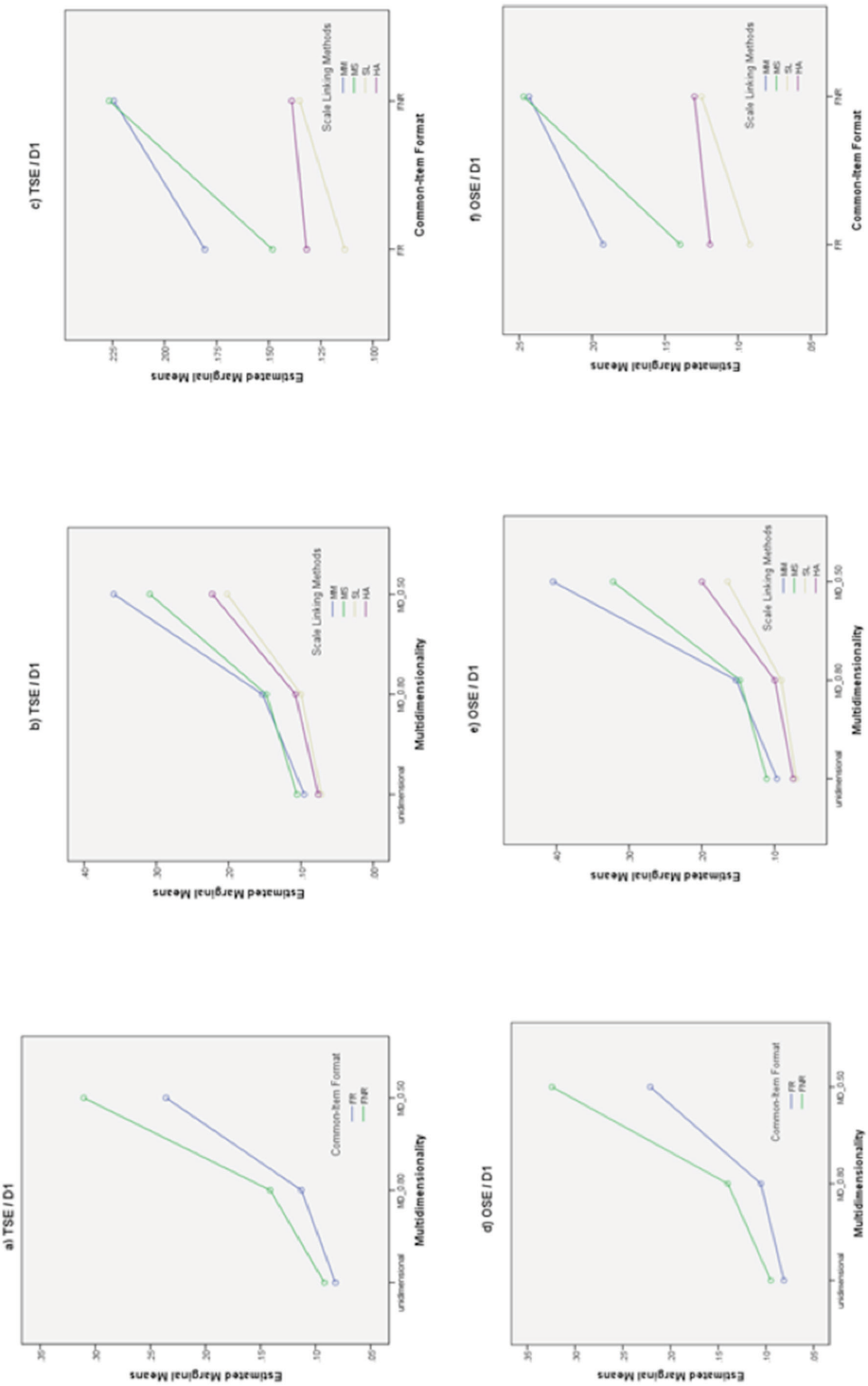

Figure 1. Factors' two-way interaction effects on FOE property under IRT true-score and observed-score equating.

Note. TSE = true-score equating, OSE = observed-score equating, FR = format representative, FNR = format non-representative, MM = mean-mean; mean-sigma, SL= Stocking-Lord, HA = Haebara.

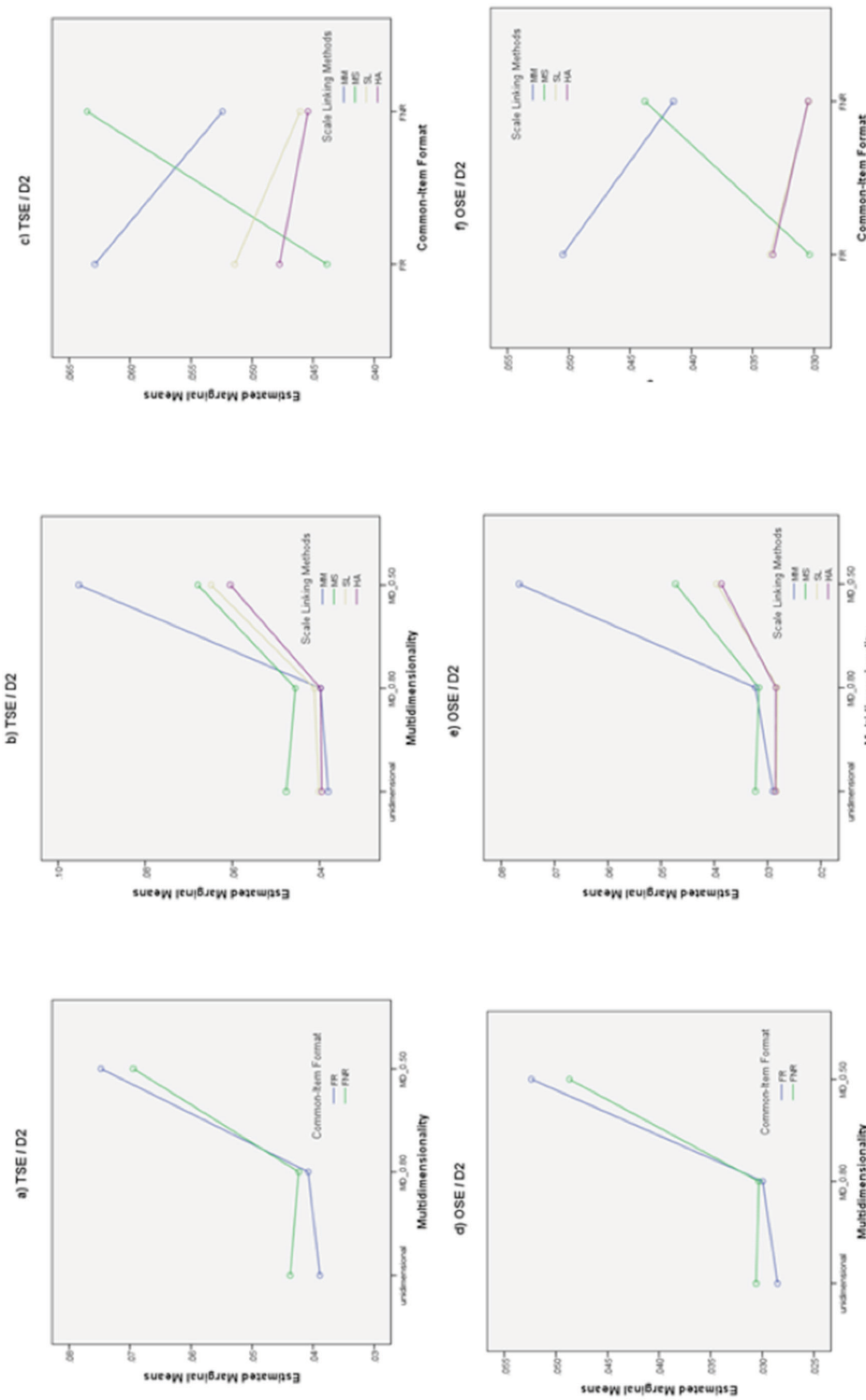

Figure 2. Factors' two-way interaction effects on SOE property under IRT true-score and observed-score equating.

Note. TSE = true-score equating, OSE = observed-score equating, FR = format representative, FNR = format non-representative, MM = mean-mean; mean-sigma, SL= Stocking-Lord, HA = Haebara.

1f showed that all four scale linking methods had the lowest D_1 values when common-item set were FR. Also, characteristic curve methods always had lower D_1 values than moment methods. With regard to SOE criteria, the *common-item format*scale linking methods* interaction had statistically a significant and small effect on TSE and OSE results [$F_{TSE}(3,4776) = 85.77$, $p < .05$, $\eta^2 = .05$; $F_{OSE}(3,4776) = 78.14$, $p < .05$, $\eta^2 = .05$]. Figure 2c and Figure 2f showed that except for the mean-sigma method, the other three methods had the smallest D_2 values when the common-item set was FNR. The MS method behaved in opposite way and had the smallest D_2 values when the common-item set was FR.

Discussion and Conclusions

In this simulation study, the impact of dimensionality, common-item set format, and different scale linking methods on mixed-format test equating was investigated based FOE and SOE properties. Findings showed that the most notable and significant effect on the equating results among the three factors was dimensionality. The TSE and OSE results best preserved FOE under unidimensional test structure and as the degree of multidimensionality increased the mean of D_1 values increased. Therefore, both equating results showed that FOE was worst preserved when the correlation between the two abilities was a value of .50. For the SOE criteria, TSE and OSE results were best preserved under a unidimensional test structure or when the multidimensionality was not severe (in other words, the correlation between abilities was .80). Again, both equating results performed poorly in terms of preserving SOE when the multidimensionality was severe. This was consistent with expectation because unidimensional IRT equating methods presuming that the assumptions of unidimensionality and local independence have been satisfied. Therefore, as Lord (1980) indicated that applying unidimensional equating methods to multidimensional data would potentially decrease equity property of scores. Our study confirmed Lord's proposal in some sense. Bolt (1999) examined the performance of the TSE method under various multidimensional test structures and he found that TSE was affected by the presence of multidimensionality. In other studies which the effects of multidimensionality on mixed-format test equating were evaluated based on various criteria, it was also found that multidimensionality had a negative effect on unidimensional test equating (Andrews, 2011; Cao, 2008). This finding should be considered by equating practitioners before equating tests with unidimensional IRT equating methods and unidimensionality assumption should be checked.

Another important finding of this study was that the common-item set had a statistically significant effect on both TSE and OSE equating results in terms of preserving FOE, but it did not have any statistically and practically significant effect on both equating results based on SOE property. Although, under unidimensional test structure format non-representative (FNR) common item set perfomed similar as format representative (FR) common-item set in terms of preserving FOE property,

as the degree of multidimensionality increased, format representativeness of the common-item set became important. This finding was consistent with other studies which investigated the impact of characteristic of common-item set. Hagge (2010) and Kirkpatrick (2005) found that if examinees found certain item formats more difficult relative to other item formats, equating mixed-format tests with only multiple-choice common items may effect equating results negatively. Cao (2008) indicated that when tests were multidimensional a FR common-item set led more accurate results. In addition, Wolf (2013) showed that especially when data were multidimensional a representative of the common-item set led lower D_1 and D_2 indices. The composition of common-item set in CINEG design is very important and the general advise is that common-item set should be representative of overall test (Kolen & Brennan, 2004). Results from our study supported this argument based on FOE criteria.

The findings also showed that when SOE property was used as criteria, common item set did not have any statistically and practically effect on both equating results. This result suggested that in terms of measurement precision including or excluding CR items to the common-item set did not make any difference. This finding agrees with Kim, Walker, and McHale's (2008) study. They found that use of CR items without trend scoring in the common-item set would lead similar results as an MC-only common-item set.

Lastly, the results of this study showed that scale linking methods had a significant and practical effect on both equating results in terms of preserving FOE and SOE properties. Characteristic curve methods generally performed better than moment methods with regard to preserve both equity properties. This result is consistent with what has been found in past studies which compared characteristic curve methods with moment methods for the dichotomous IRT models (Hanson & Beguin, 2002; Kim & Cohen, 1992; Ogasawara, 2002) and mixture IRT models (Kim, 2004; Kim & Lee, 2004; Kim & Lee, 2006). The characteristic curve methods require an iterative multivariate search procedure but moment methods require simple summary statistics (Kim, 2004). As Ogasawara (2002) indicated, item characteristic curves could be estimated accurately, though item parameters were not estimated very precisely. Therefore, IRT characteristic curve linking methods which using item/test response functions give more stable results than moment methods (Ogasawara, 2001). As found in previous studies (Kim, 2004; Kim & Kolen, 2006), the results showed that the two characteristic curve scale transformation methods performed similar and they were more robust to presence of multidimensionality.

Overall, the result of this study suggested that FOE and SOE properties were best preserved when tests were unidimensional, common-item set was format representative, and the test characteristic linking methods were used. This study is limited with simulated data, although we tried to mimic the real data situation simulations cannot capture all features of the real testing environment. Therefore, to be able to generalize

conclusions of this study to practical situations, it is necessary to repeat this study using real data. Also, this study is limited only three factors and equity evaluation criteria. Future studies should be done with other factors such as length of common-item set, different IRT models, sample size, form differences and different evaluation criteria.

References

Andrews, B. J. (2011). *Assessing first- and second-order equity for the common item nonequivalent groups design using multidimensional IRT* (Doctoral dissertation). Available from ProQuest Dissertation and Theses database. (UMI No. 3473138)

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 147–162.

Balch, J. (1964). The influence of the evaluating instrument on students' learning. *American Educational Research Journal, 1*(3), 169–182.

Bastari, B. (2000). *Linking multiple-choice and constructed-response items a common proficiency scale* (Doctoral dissertation). Available from ProQuest Dissertation and Theses database. (UMI No. 9960735)

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education*, *12*(4), 383–407. Retrieved from http://dx.doi.org/10.1207/S15324818AME1204_4

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York, NY: Academic.

Brennan, R. L. (2010). *First-order and second-order equity in equating* (Report No. 30). Iowa City: Center for Advanced Studies in Measurement and Assessment.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common item sets* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3341415)

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). New York, NY: Springer.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149. Retrieved from https://www.jstage.jst.go.jp/article/psycholres1954/22/3/22_3_144/_pdf

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3422144)

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(3), 3–24. http://dx.doi.org/10.1177/0146621602026001001

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*, 195–240. http://dx.doi.org/10.1207/s15324818ame0603_3

He, Y. (2011). *Evaluating equating properties for mixed-format tests* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3461151)

Kim, S. (2004). *Unidimensional IRT scale linking procedures for mixed-format tests and their robustness to multidimensionality.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3129309)

Kim, S-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51–66.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models* [Computer software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

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. doi: 10.1207/s15324818ame1904_7

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests*. ACT research report series. Iowa City, IA. (Eric Document ED484785).

Kim, S., & Lee. W. C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53–76. http://dx.doi.org/10.1111/j.1745-3984.2006.00004.x

Kim, S., Walker, M. E., & McHale, F. (2008). *Equating of mixed-format tests in large-scale assessments*. Technical Report (RR-08-26). Princeton, NJ: Educational Testing Service.

Kirkpatrick, R. K. (2005). *The effects of item format in common item equating* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3184724)

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational and Behavioral Statistics, 9*(1), 25–44. http://dx.doi.org/10.3102/10769986009001025

Kolen, M. J. (2004a). *POLYEQUATE windows console version* [Computer software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

Kolen, M. J. (2004b). *POLYCSEM windows console version* [Computer software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

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

Livingston, S. A. (2009). *Constructed-Response test questions: Why we use them; How we score them* (R & D Connections, No. 11). Princeton, NJ: Educational Testing Service. (Eric Document ED507802).

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139–160. http://dx.doi.org/10.1111/j.1745-3984.1977.tb00033.x

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chesnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218. http://dx.doi.org/10.1207/s15326985ep3404_2

Messick, (1963). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61–73). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.) *Test equating* (pp. 169–191). New York, NY: Academic.

Mullis, I. V. S., & Martin, M. O. (2011). *TIMSS 2011 item writing guidelines*. Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. http://dx.doi.org/10.1177/014662169201600206

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1* [Computer software]. Chicago, IL: Scientific Software International, Inc.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25*(1), 53–67. http://dx.doi.org/10.1177/01466216010251004

Ogasawara, H. (2002). Stable response functions with unstable item parameter estimates. *Applied Psychological Measurement*, *26*(3), 239–254. http://dx.doi.org/10.1177/0146621602026003001

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York, NY: Springer.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249–275. http://dx.doi.org/10.1111/j.1745-3984.2007.00037.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210. http://dx.doi.org/10.1177/014662168300700208

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29*(6), 418–432. http://dx.doi.org/10.1177/0146621606280071

Traub, R. E. (1993). On the equivalence of traits assessed by multiple-choice and constructed response tests. In R. E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29–44). Hillsdale, NJ: Erlbaum.

Wang, W. (2013). *Mixed-format test score equating: Effect of item-type multidimensionality, length and composition of common-item set, and group ability difference* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3608502)

Wolf, R. (2013). *Assessing the impact of characteristics of the test, common items, and examinees on the preservation of equity properties in mixed-format test equating* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3585536)

Yao, L. (2003). *SimuMIRT* [Computer software]. Monterey, CA: Defense Manpower Data Center.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469–492. http://dx.doi.org/10.1177/0146621605284537