ACT Research Report Series 2010-1

Evaluating the Effects of Differences in Group Abilities on the Tucker and the Levine Observed-Score Methods for Common-Item Nonequivalent Groups Equating

Hanwei Chen

Zhongmin Cui

Rongchun Zhu

Xiaohong Gao

ACT

July 2010

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

Evaluating the Effects of Differences in Group Abilities on the Tucker and the Levine Observed-Score Methods for Common-Item Nonequivalent Groups Equating

Hanwei Chen
Zhongmin Cui
Rongchun Zhu
Xiaohong Gao

Abstract

The most critical feature of a common-item nonequivalent groups equating design is that the average score difference between the new and old groups can be accurately decomposed into a group ability difference and a form difficulty difference. Two widely used observed-score linear equating methods, the Tucker and the Levine observed-score methods, have different statistical assumptions when decomposing the score difference. Variation in the decomposition of group ability and form difficulty differences can affect the equating results.

This study confirmed previous findings in the literature that when form and group differences are small, both equating methods produce similar results. When the group ability difference is large, however, the Levine observed-score method produces more accurate equating results than the Tucker method. The results indicated that the Levine observed-score method not only decomposes form and group differences more accurately, but also yields smaller unweighted absolute equating differences and average weighted root mean square differences. This study showed that the Levine observed-score method is also robust to the form difference.

Evaluating the Effects of Differences in Group Abilities on the Tucker and the Levine Observed-Score Methods for Common-Item Nonequivalent Groups Equating

Introduction

A common-item nonequivalent groups equating design is often used in many testing programs because of its flexibility in data collection. Important features of this design include: (1) each of the two examinee groups (new and old) is only required to take one alternative form of the test; (2) a set of common items is embedded in both the new and old forms, which links the two forms of the test; and (3) the common-item set should be viewed as a short version of the full-length test, which requires similar content and statistical specifications (including difficulty). Among the applicable equating methods under the common-item nonequivalent groups design, two observed-score linear equating methods are of particular interest: the Tucker and the Levine observed-score equating methods.

Because each examinee only takes one alternative form of the test, strong statistical assumptions are necessary in establishing the linear equating function for the new and old forms. Two statistical assumptions about the observed scores are made for the Tucker equating method: linear regression and conditional variances. The linear regression assumption indicates that the regression of the total scores on the common-item scores is the same for both the new and old populations. The conditional variances assumption requires that the conditional variances of the total scores given the common-item scores are the same in both populations.

On the other hand, three statistical assumptions are made for the Levine observed-score equating method: correlational assumptions, linear regression assumptions, and error variance assumptions. The correlational assumptions specify that the true scores for the forms and the common-items are perfectly correlated in the new and old populations. The linear regression assumptions mean that the regressions of the true scores for the new form (or old form) on the

true scores for the common-items are the same for both the new and old populations. Furthermore, the error variance assumption means that the measurement error variances for the new form, old form, or common-items are the same for both the new and old populations (see Kolen and Brennan, 2004, pp. 105-117 for details).

When all the assumptions are satisfied, research has indicated that both equating methods will produce the same results (Kolen, 1990; von Davier, 2008). von Davier (2008) indicates that the Tucker and the Levine observed-score methods can produce theoretically the same equating results when the populations are the same and all assumptions for both equating methods are satisfied. Kolen (1990) suggests that if the two populations are similar in ability and the common-item scores are highly correlated with the total scores on the two forms of the test in a common-item nonequivalent groups design, all equating methods tend to produce the same results.

Further, when comparing the Tucker and the Levine observed-score equating methods both empirically and theoretically, Kolen and Brennan (2004, pp.128-129) suggested that the equating decisions favor (a) the Tucker method, when both examinee groups are similar in ability; (b) the Levine observed-score method, when the examinee groups are dissimilar in ability; or (c) not conducting equating, if the examinee groups are very different in ability or the forms are too much dissimilar in difficulty. However, in practical situations, both form and group differences can exist in an equating and the magnitudes of the differences may vary. Therefore, under the common-item nonequivalent groups design, the interaction between examinee group difference and form difference is crucial to the equating results based on the different equating methods.

Under the common-item nonequivalent groups equating, the observed total score mean differences are due to the confounded effects of form and group mean differences. They can be mathematically decomposed into the sum of form and group mean differences:

$$\mu_1(X) - \mu_2(Y) = \{\mu_1(X) - \mu_2(Y) - \gamma_2[\mu_1(V) - \mu_2(V)]\} + \{\gamma_2[\mu_1(V) - \mu_2(V)]\} \quad (1)$$

where $\mu_1(X)$ denotes the observed total score mean for new form in new group, $\mu_2(Y)$ denotes the observed total score mean for old form in old group, $\mu_1(V)$ denotes the observed common-item score in new group, $\mu_2(V)$ denotes the observed common-item score in old group, and γ_2 denotes the expansion factor. Note that the first brackets represent the form difference (or FD) for the new group and the second brackets represent the group difference (or GD) on the old form scale. Equation (1) applies to both the Tucker and the Levine observed-score methods under the condition that the synthetic population is the new group with weight equal to 1.

The group difference on the common-items is computed by multiplying an "expansion factor" (or γ_2). The expansion factors are:

$$\gamma_2(Tucker) = \alpha_2(Y \mid V) = \frac{\sigma_2(Y,V)}{\sigma_2^2(V)} \quad (2)$$

$$\gamma_2(Levine) = \frac{1}{\alpha_2(V \mid Y)} = \frac{\sigma_2^2(Y)}{\sigma_2(Y,V)} \quad (3)$$

where $\alpha_2(Y \mid V)$ denotes the regression slope for Y given V for the old group, $\alpha_2(V \mid Y)$ the regression slope for V given Y, $\sigma_2(Y,V)$ the covariance of Y and V for the old group, $\sigma_2^2(Y)$ the variance of old form for the old group, and $\sigma_2^2(V)$ the variance of V for the old group. The different expansion factors are due to the different statistical assumptions of the equating methods. Kolen and Brennan (2004) indicated that because the Levine expansion factors are usually larger than the Tucker's, population differences under the Levine assumptions are greater

than under the Tucker assumptions. They suggested that the Levine observed-score method is more appropriate than the Tucker method when examinee populations are dissimilar.

It should be noted that the expansion factor has a role on not only the decompositions of form and group differences but also the slopes of determining linear equating functions:

$$l_Y(x) = \frac{\sqrt{\sigma_2^2(Y) + \gamma_2^2[\sigma_1^2(V) - \sigma_2^2(V)]}}{\sigma_1(X)}[x - \mu_1(X)] + \mu_2(Y) \qquad (4)$$

where $l_Y(x)$ denotes the linear equating function for the new form on the old form scale, $\sigma_1^2(V)$ and $\sigma_2^2(V)$ the variances of common-item for the new and old group respectively, and $\sigma_1(X)$ the standard deviation of new form for the new group.

As decomposing the observed total score mean differences into form and group mean differences is taken as the intermediate process and plays an important role in the common-item nonequivalent groups equating, it is critical to understand how the equating methods decompose the observed total scores and what impact these methods have under various form and group difference conditions.

When different equating methods are implemented in the common-item nonequivalent groups design, the equating results should be in line with the degree of similarity in group abilities and form difficulties (e.g., forms that differ in difficulty would have conversions that are more disparate than forms that are similar in difficulty). Many studies (e.g., von Davier, 2008; Wang, Lee, Brennan, & Kolen, 2006) have used simulation techniques to manipulate either GD or FD separately. Since both group difference and form difference may interact, the goal of this study is to comprehensively investigate the impact of combining group difference and form difference (with more weight on group difference) on equating results for the Tucker and the Levine equating methods.

Including similar and dissimilar conditions for both group and form differences, the current study conducted 2 X 2 analyses based on the Tucker and the Levine observed-score equating methods. Four variations can be categorized: similar in both forms and ability, similar in forms and dissimilar in ability, dissimilar in forms and similar in ability, and dissimilar in both forms and ability.

Combining the four categories and the three equating decision suggestions from Kolen and Brennan (2004) listed above, three research questions are investigated in the current study: (1) When the test forms and examinee groups are similar, do both equating methods produce the same results? (2) When the test forms are similar in difficulty, which of the equating methods is more robust in dealing with differences in group ability? and (3) When the test forms are dissimilar in difficulty, which of the equating methods is more robust in dealing with differences in group ability?

Method

Data

Test data from a nationally administered 30-item mathematics test were used in this study. Examinee groups with similar educational backgrounds took four test forms (denoted as Forms A, B, C, and D) in different administrations. Forms A and B share 10 carefully selected common-items which are similar to the test as a whole in terms of content and statistical properties. The dichotomous item responses on both forms were calibrated using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) with a 3PL IRT model (Lord, 1980). The Stocking and Lord Method (Stocking & Lord, 1983) was applied to place item parameters of Forms A and B on the same scale. The average IRT b parameter values on Forms A and B were 0.097 and 0.004 respectively, thus Forms A and B were used in the condition of similar form difficulty.

Similarly, Forms C and D also have 10 items in common, and item parameters of both forms were placed on the same scale using the Stocking and Lord Method. The average IRT b parameter values on Forms C and D were 0.125 and -0.263 respectively, thus Forms C and D were used in the condition of dissimilar form difficulty.

The empirical item parameter values were assumed to be the true parameter values and were used as the basis to simulate item responses. To simulate the condition of group difference, five 2000-examinee groups with mean abilities ranging from -0.5 to 0.5 (i.e., -0.5, -0.25, 0, 0.25, 0.5) were sampled from a normal distribution with unit standard deviation and were denoted as L1, L, Mid, H, and H1 respectively. These groups of examinees took Forms A and C (old forms). Similarly, another five groups of examinees took Forms B and D (new forms). Pairing the groups who took new forms and corresponding old forms yielded different conditions on group difference which are described in the following section. WinGen2 (Han, 2007) was used to sample examinees and simulate item responses. The observed score distributions were used to conduct equating.

Equating Conditions

Both the Tucker and the Levine observed-score equating methods were conducted for all equating conditions. The synthetic population weight was set to 1 for the new group, which means that the new group is considered to be the sole synthetic group. Factors of investigation are described below:

1. Equating methods: Tucker and the Levine observed-score equating.
2. Form difficulty differences: 0.1 (similar) and 0.4 (dissimilar) measured by the mean difference of *b* parameter for the new and old forms.
3. Group differences: from 0 to 1 with an increment of 0.25 on the theta scale.

The group ability difference was measured by the mean difference between the two normal distributions from which random samples of thetas were drawn. In the current study the new group is considered less able than the old group, because the magnitudes of group differences are the same except for opposite signs when the new group is considered more able than the old group. Thus, fifteen group ability combinations are computed for each of the two form difficulty conditions (similar and dissimilar) including ***0.00*** (same ability for new and old groups: *SH1, SH, SM, SL, and SL1*), ***0.25*** (0.25 ability difference for new and old groups: *HH1, MH, LM, and L1L*), ***0.50*** (0.5 ability difference for new and old groups: *MH1, LH, and L1M*), ***0.75*** (0.75 ability difference for the new and old groups: *LH1 and L1H*), and ***1.00*** (1 ability difference for new and old groups: *L1H1*). A total of 30 (2 form conditions X 15 ability combinations) equatings were conducted separately using Tucker and Levine observed-score equating methods. Table 1 shows the group difference conditions investigated in this study.

TABLE 1

Equating Conditions and Notations

		Old Form				
		High1: N(0.5, 1)	High: N (0.25, 1)	Mid: N(0, 1)	Low: N(-0.25, 1)	Low1: N(-0.5, 1)
New Form	High1: N(0.5, 1)	SH1				
	High: N(0.25, 1)	HH1	SH			
	Mid: N(0, 1)	MH1	MH	SM		
	Low: N(-0.25, 1)	LH1	LH	LM	SL	
	Low1: N(-0.5, 1)	L1H1	L1H	L1M	L1L	SL1

Note: (0.00)_Diff includes SH1, SH, SM, SL and SL1
(0.25)_Diff includes HH1, MH, LM and L1L
(0.50)_Diff includes MH1, LH and L1M
(0.75)_Diff includes LH1 and L1H
(1.00)_Diff includes L1H1

Evaluating Indices

It is a common practice that the selected equating methods are evaluated against a true equating function. In this study, each ability group was considered as a subgroup from the target population. Thus, the five examinee groups taking the new forms (or the old forms) were combined and used to represent the examinee population. The criterion equating group is denoted as Mix10K. That is, two 10,000-examinee groups were used to conduct the equating under the two form difficulty conditions. The equating relationship of the examinee population taking the new and old forms (i.e., the Mix10K) was adopted as the criterion. The equipercentile equating method is used to define the criterion equating relationship. Another way of defining a target population is to draw a large number of examinees (e.g., 100,000) from a standard normal

distribution. However, this consideration was not adopted here because it fails to represent examinee groups with different abilities.

The decompositions of form difference and group difference were evaluated for their accuracy under each method in terms of their capability of separating the two differences. Under the same condition of form difficulty (similar or dissimilar), the decomposed form differences are expected to be the same regardless of the changing differences in group abilities. That is, the consistency of decomposing form difference is to be compared between the equating methods when group differences change. Likewise, when the same pair of groups (e.g., LM or MH) are used, the decomposed group differences are expected to be the same regardless of the changing difference in form difficulty (i.e., similar or dissimilar in form difficulty). The accuracy of decomposing group difference is to be compared between the equating methods when form differences change.

Different equating conditions may lead to different equating results which can be quantified and evaluated through the produced unrounded scale score conversions. The old form conversion was used in such a way that both unrounded and reported scale scores were identical to the associated raw scores. Since the equipercentile equating relationship between the Mix10K groups was adopted as the criterion equating, the average unrounded scale score differences between each of the equating conditions and the criterion equating were compared. The indices used for evaluating the equating results were: (1) the average unweighted absolute equating difference (UAED); and (2) the average weighted root mean square difference (RMSD) at each score point. Computation formula was listed below:

$$UAED_j = \frac{\sum_{i=0}^{30} \left| y_j(x_i) - y_{cri}(x_i) \right|}{31}, j = 1,...,15 \qquad (5)$$

where x_i denotes the raw score; j denotes one of the fifteen equating conditions (i.e., SH1, SH, SM, SL, SL1, HH1, MH1, LH1, L1H1, MH, LH, L1H, LM, L1M, and L1L); y_{cri} denotes the unrounded scale score conversion for the criterion; and $y_j(x_i)$ denotes the unrounded scale score for the corresponding raw score x_i under equating condition j.

The root mean square difference (RMSD) was used to compute the average of the weighted squared difference between the varying equating conditions and the criterion equating (see Dorans & Holland, 2000):

$$RMSD_j = \frac{\sqrt{\sum_{i=0}^{30} w_i [y_j(x_i) - y_{cri}(x_i)]^2}}{\sigma_{cri}}, j = 1,...,15 \qquad (6)$$

where w_i denotes the proportion of new form sample with raw score x_i, σ_{cri} denotes the standard deviation of the old criterion group, and the other notations are the same as above.

Results

Descriptive Statistics for Forms and Samples

Tables 2 and 3 present average item parameter estimates for Forms A and B (similar difficulty) as well as for Forms C and D (dissimilar difficulty). Table 4 shows the means and standard deviations for values of theta, total raw score, common-item score, and non-common item score for each ability group under different form difficulty conditions. The results indicate that the ability difference decreases by 0.25 on the theta scale as the total raw score decreases by one point when form difficulty is similar (S-new and S-old). This one-point difference can be attributed to about 0.3 point and 0.7 point differences for common-item and non-common item scores, respectively.

TABLE 2

Descriptive Statistics for Item Parameters for Similar Form Difficulties

Parameter	n	Mean	SD	Skewness	Kurtosis
Form_A (new)					
a	30	1.107	0.438	0.230	-0.456
b	30	0.097	1.696	-0.449	-0.853
c	30	0.183	0.132	1.172	1.518
Common-Item Set					
a	10	1.152	0.339	0.074	1.201
b	10	0.585	1.454	-0.624	-0.383
c	10	0.214	0.121	1.357	3.819
Form_B (old)					
a	30	1.132	0.278	0.714	0.527
b	30	0.004	1.492	-0.125	-1.157
c	30	0.133	0.099	1.183	2.770
Common-Item Set					
a	10	1.184	0.237	1.264	1.385
b	10	0.493	1.402	-0.558	-0.344
c	10	0.170	0.077	-0.738	1.120

TABLE 3

Descriptive Statistics for Item Parameters for Dissimilar Form Difficulties

Parameter	n	Mean	SD	Skewness	Kurtosis
Form_C (new)					
a	30	1.068	0.371	-0.072	-1.488
b	30	0.125	1.820	-0.706	-0.555
c	30	0.146	0.085	0.652	-0.563
Common-Item Set					
a	10	1.173	0.418	-0.735	-0.321
b	10	0.536	1.198	-0.335	0.287
c	10	0.144	0.079	0.287	-0.884
Form_D (old)					
a	30	1.144	0.416	0.390	-0.821
b	30	-0.263	1.592	-0.428	-0.748
c	30	0.121	0.086	0.609	-0.970
Common-Item Set					
a	10	1.174	0.426	-0.655	-1.051
b	10	0.511	1.234	-0.069	-0.619
c	10	0.142	0.083	0.368	-0.675

TABLE 4

Means and Standard Deviations for Theta, Total Raw Score, Common-Item Score, and Non Common-Item Score – Varying Ability Groups

		Mean				SD			
		S-new Form A	S-old Form B	DS-new Form C	DS-old Form D	S-new Form A	S-old Form B	DS-new Form C	DS-old Form D
Theta	HI	0.508	0.482	0.508	0.482	0.993	1.009	0.993	1.009
	H	0.255	0.279	0.255	0.279	0.995	0.999	0.995	0.999
	Mid	0.005	0.031	0.005	0.031	0.997	1.002	0.997	1.002
	L	-0.243	-0.260	-0.243	-0.260	0.997	0.980	0.997	0.980
	LI	-0.490	-0.474	-0.490	-0.474	1.014	1.001	1.014	1.001
	Ref	0.008	0.012	0.008	0.012	1.058	1.056	1.058	1.056
Raw total	HI	17.907	19.052	17.878	19.504	4.681	4.695	4.650	4.887
	H	17.052	18.072	16.852	18.497	4.545	4.657	4.538	4.917
	Mid	16.025	17.074	15.931	17.495	4.424	4.744	4.421	4.836
	L	14.988	15.978	14.975	16.208	4.363	4.596	4.310	4.652
	LI	14.146	15.151	14.124	15.358	4.225	4.535	4.200	4.604
	Ref	16.084	17.028	15.955	16.084	4.621	4.857	4.657	4.622
CI	HI	5.637	5.830	5.513	5.591	1.984	1.999	2.206	2.177
	H	5.316	5.501	5.106	5.198	1.909	1.972	2.147	2.137
	Mid	4.920	5.204	4.724	4.834	1.892	1.998	2.044	2.117
	L	4.649	4.835	4.316	4.397	1.841	1.938	2.012	1.945
	LI	4.351	4.502	3.999	4.104	1.856	1.889	1.906	2.037
	Ref	5.002	5.140	4.727	4.759	1.942	2.008	2.145	2.074
Non CI	HI	12.270	13.222	12.366	13.914	3.229	3.187	2.973	3.215
	H	11.736	12.572	11.746	13.299	3.177	3.243	2.947	3.301
	Mid	11.106	11.870	11.207	12.662	3.129	3.308	2.964	3.234
	L	10.340	11.143	10.659	11.811	3.118	3.276	2.879	3.263
	LI	9.795	10.650	10.126	11.255	2.992	3.233	2.905	3.175
	Ref	11.083	11.888	11.227	11.325	3.233	3.380	3.064	3.097

Note: CI = common-item; S-old = similar old form; S-new = similar new form; DS-old = dissimilar old form; DS-new = dissimilar new form.

Table 5 shows the decompositions of total raw score differences to common-item and non-common item score differences for different equating conditions (form difficulty and varying group abilities). The results indicate that there is about a 0.5 point total raw score difference (e.g., for Zero difference, the average changes from -1.042 to -1.460) when equating conditions of form difference are changed from similar (i.e., Form A to Form B) to dissimilar (i.e., Form C to Form D).

TABLE 5

Total Raw Score Decomposition to Common-Item and Non-Common-Item Scores

Difference		Form A to Form B			Form C to Form D		
		Raw	CI	NCI	Raw	CI	NCI
Zero	SH1	-1.145	-0.193	-0.952	-1.626	-0.078	-1.548
	SH	-1.020	-0.185	-0.836	-1.645	-0.091	-1.553
	SM	-1.049	-0.284	-0.765	-1.565	-0.110	-1.455
	SL	-0.990	-0.186	-0.804	-1.233	-0.080	-1.153
	SL1	-1.006	-0.151	-0.855	-1.234	-0.105	-1.129
	avg	**-1.042**	-0.200	-0.842	**-1.460**	-0.093	-1.368
0.25	HH1	-2.000	-0.514	-1.486	-2.652	-0.485	-2.168
	MH	-2.047	-0.581	-1.466	-2.566	-0.474	-2.093
	LM	-2.086	-0.555	-1.531	-2.521	-0.518	-2.003
	L1L	-1.832	-0.484	-1.348	-2.084	-0.398	-1.686
	avg	**-1.991**	-0.533	-1.458	**-2.456**	-0.468	-1.987
0.5	MH1	-3.027	-0.910	-2.117	-3.574	-0.866	-2.707
	LH	-3.084	-0.852	-2.232	-3.522	-0.882	-2.641
	L1M	-2.928	-0.853	-2.075	-3.371	-0.835	-2.536
	avg	**-3.013**	-0.872	-2.141	**-3.489**	-0.861	-2.628
0.75	LH1	-4.064	-1.181	-2.883	-4.530	-1.275	-3.255
	L1H	-3.927	-1.150	-2.777	-4.373	-1.199	-3.174
	avg	**-3.995**	-1.166	-2.830	**-4.451**	-1.237	-3.214
One	L1H1	**-4.906**	-1.479	-3.427	**-5.380**	-1.592	-3.788
Ref	Mix10K	-0.944	-0.139	-0.805	-1.447	-0.082	-1.366

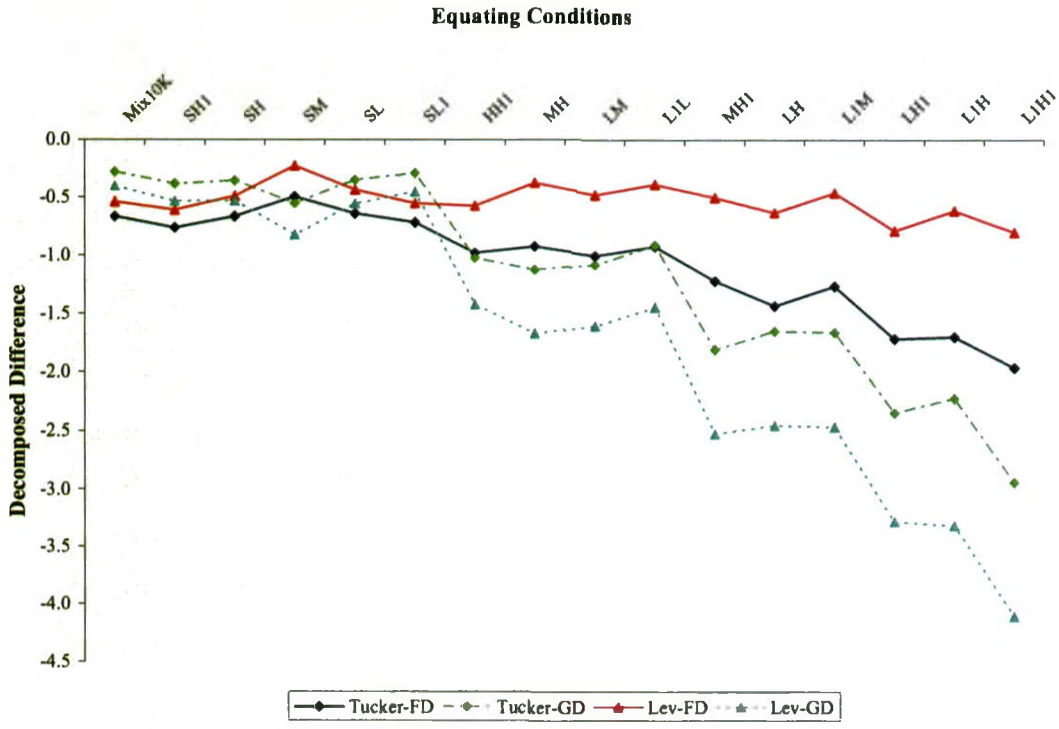
Decompositions of Form and Group Differences

Consistencies of the decompositions are evaluated for each equating method under different conditions. Under the same form difficulty condition, it is expected that the decomposed form differences are consistent regardless of the difference in group abilities.

Figures 1 and 2 present the decomposed form and group differences from each equating method when form difficulty is similar or dissimilar. As expected, when the groups are very similar (i.e., SH1, SH, SM, SL, SL1) both methods yield similar decompositions of the group and form differences. Both methods captured the form difference when they are dissimilar

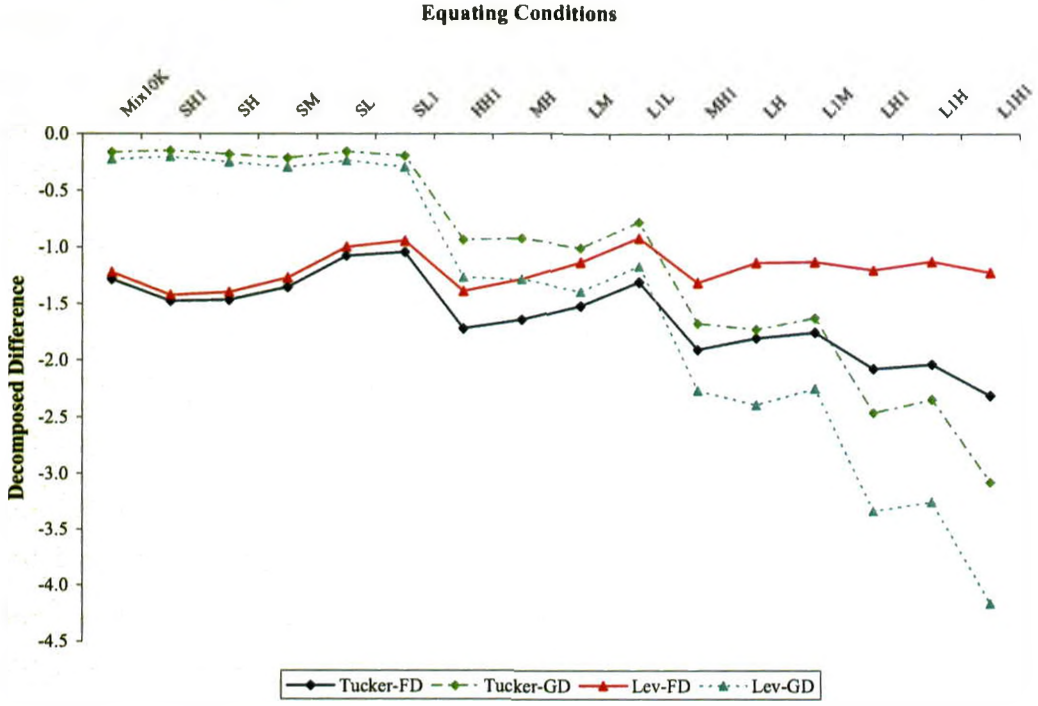
(Figure 2). When the group difference increases (e.g., greater than 0.25 SD including MH1, LH, L1M, LH1, L1H, and L1H1), the decomposed values of the group difference also increase from both methods. However, the magnitudes of the decomposed group difference vary between the two methods, with the Levine observed-score method yielding larger group differences than the Tucker method. The patterns are consistent for the similar (Figure 1) and dissimilar (Figure 2) form conditions. Additionally, the decompositions of form differences from the Levine observed-score method are consistent regardless of the change in group differences. The decompositions of form differences from the Tucker method, however, become larger as the difference in groups increases. This finding suggests that the Levine observed-score method is more robust than the Tucker method in the decomposition of form differences when group differences change.

FIGURE 1. Decomposition of group and form differences for similar forms: Forms A and B.



Note: FD = Form Difference; GD= Group Difference

FIGURE 2. Decomposition of group and form differences for dissimilar forms: Forms C and D.



Note: FD = Form Difference; GD= Group Difference

Likewise, in the same group ability condition, it is expected that the decomposed group differences are consistent regardless of the change in form difficulties. For the same pair of test forms having similar and dissimilar form difficulties, Figure 3 presents the decomposed group differences of the equating methods for all equating conditions. The result shows that, when the form difference changes from similar to dissimilar, the group difference does not change much for both methods. Additionally, the Levine observed-score method yields slightly larger group differences than the Tucker method in most of the equating conditions. In general, the Levine observed-score method appears to be better than the Tucker method in the sense of revealing the relative magnitude of the decomposed differences, especially when group differences are dissimilar. This finding is consistent with the literature. Figure 4 presents the decomposed form differences with the changes of form similarity. In general, the form differences appear larger

under the Tucker method than the Levine observed-score method, which is also consistent with the literature. Furthermore, the decomposed form differences for Tucker method increase slightly as the group variation increases even for the similar forms.

FIGURE 3. Decomposition of group difference for form difficulty conditions: Similar vs. Dissimilar

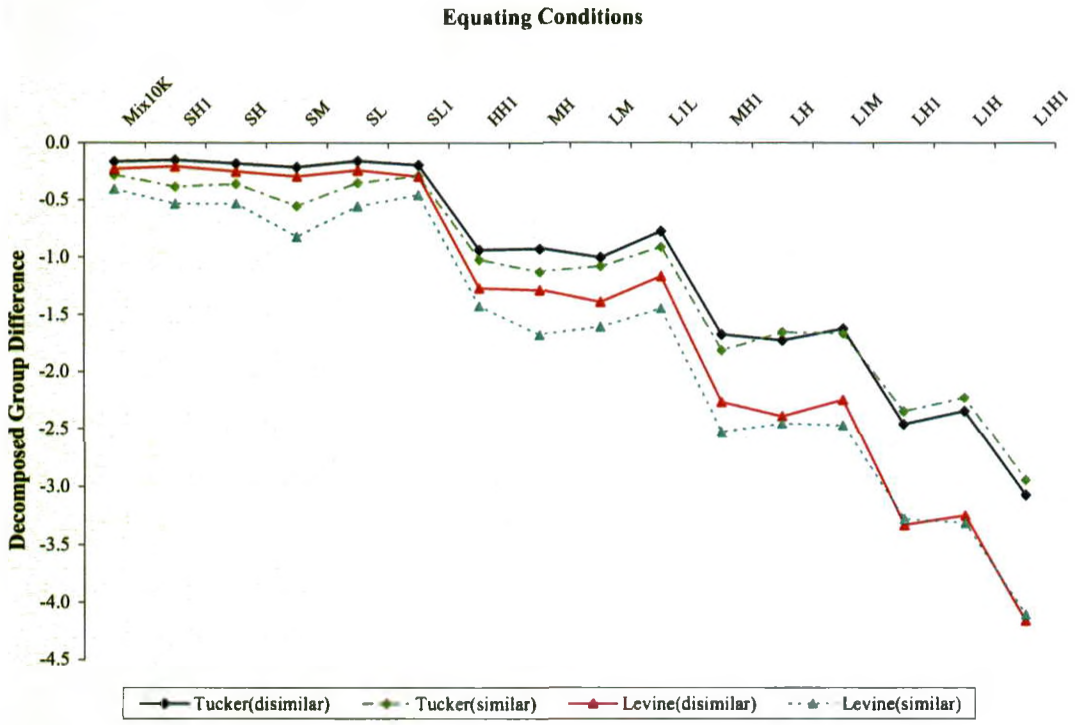
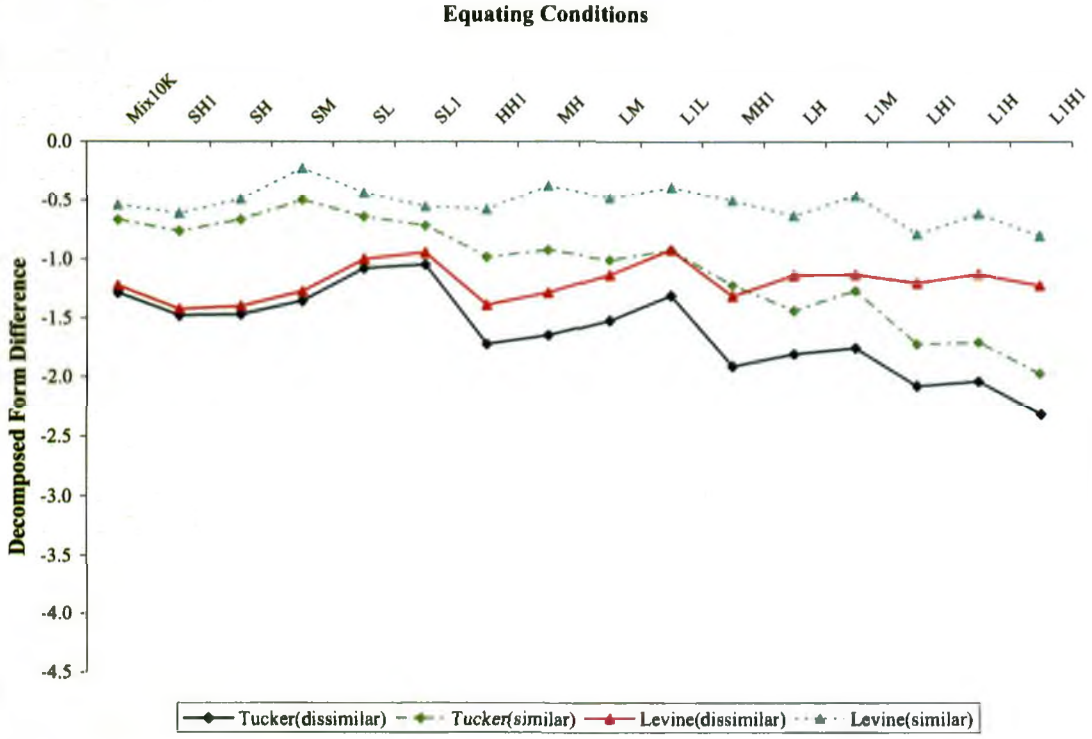


FIGURE 4. Decomposition of form difference for form difficulty conditions: Similar vs. Dissimilar



Equated Unrounded Scale Scores

Tables 6 to 9 present unrounded scale score means for the new groups (High1, High, Mid, Low, and Low1) using the Tucker and the Levine observed-score equating methods when the new and old forms are either similar (Tables 6 and 7) or dissimilar (Tables 8 and 9) in difficulty and the new and old group abilities are either similar or different. It was expected that the unrounded scale score means for the same new group should be consistent across the different conditions. However, the results indicate that the unrounded scale score means depend on the similarities between the new and old group abilities and new and old form difficulties under the Tucker equating.

TABLE 6

Unrounded Scale Score Means for Similar Forms –Tucker

Tucker	Similar Forms (Form A to Form B)				
New	High1	High	Mid	Low	Low1
High1	**18.669**				
High	18.031	**17.714**			
Mid	17.242	16.947	**16.520**		
Low	16.704	16.422	15.992	**15.627**	
Low1	16.112	15.845	15.412	15.066	**14.862**

TABLE 7

Unrounded Scale Score Means for Similar Forms –Levine

Levine	Similar Forms (Form A to Form B)				
New	High1	High	Mid	Low	Low1
High1	**18.517**				
High	17.626	**17.541**			
Mid	16.526	16.399	**16.251**		
Low	15.774	15.618	15.467	**15.422**	
Low1	14.947	14.760	14.604	14.532	**14.697**

TABLE 8

Unrounded Scale Score Means for Dissimilar Forms –Tucker

Tucker	Dissimilar Forms (Form C to Form D)				
New	High1	High	Mid	Low	Low1
High1	**19.354**				
High	18.569	**18.318**			
Mid	17.833	17.572	**17.282**		
Low	17.046	16.775	16.490	**16.050**	
Low1	16.433	16.155	15.873	15.430	**15.165**

TABLE 9

Unrounded Scale Score Means for Dissimilar Forms –Levine

Levine	Dissimilar Forms (Form C to Form D)				
New	High1	High	Mid	Low	Low1
High1	**19.300**				
High	18.238	**18.249**			
Mid	17.240	17.214	**17.201**		
Low	16.175	16.109	16.105	**15.972**	
Low1	15.345	15.248	15.252	15.042	**15.066**

For example, the new group, Low1, is equated to more able old groups including High1, High, Mid, Low, and to similar Low1. The unrounded scale score means for Low1 are compared across the different old groups and between the equating methods. In Tables 6 and 7, when the test forms are similar in difficulty, the Tucker method produces larger unrounded scale score means than the Levine observed-score method, especially when the new and old group differences are large. The unrounded scale score means from the Tucker method are 16.112, 15.845, 15.412, 15.066, and 14.862 for the old groups of High1, High, Mid, Low, and Low1 respectively; whereas, the Levine observed method yields 14.947, 14.760, 14.604, 14.532, and 14.697 corresponding to the unrounded scale score means. The results indicate that the unrounded scale score means from the Levine observed-score equating are more consistent than those from the Tucker equating across the levels of group ability difference between the new and old groups. The findings suggest that, when the examinee groups are dissimilar in abilities, the Tucker method tends to count the group difference as form difference. As a result, the more able group is disadvantaged while the less able group is advantaged (from the Tables 6 and 8). Depending on the direction of equating (i.e., less able new group to more able groups, or vice-versa), the equating score may be inflated or deflated by the Tucker method. The same pattern can be found for the test forms with dissimilar difficulties (see Tables 8 and 9). These results

suggest that the differences in group abilities between the new and old groups do not affect the equated scores under the Levine observed-score equating but do affect the equating results under the Tucker equating.

Figures 5 and 6 present the unrounded equated score differences between equatings from different samples (High1, High, Mid, Low, and Low1) and the criterion equating when the new and old forms and groups are similar. The results are similar between the two equating methods: (1) the equated scores are not very different from the criterion equating in the middle score range; (2) the equated scores at the lower end of the scale are more similar to the criterion equating when using the Low and Low1 samples than the high ability samples; (3) the equated scores at the higher end of the scale are more similar to the criterion equating when using the High and High1 samples than the low ability sample; and (4) the Tucker method performs slightly better than the Levine observed-score method except at the high end of the score scale when the middle ability samples are used.

FIGURE 5. Differences of unrounded equated scores between various equatings and criterion equating: Form Similar & Ability Similar – Tucker

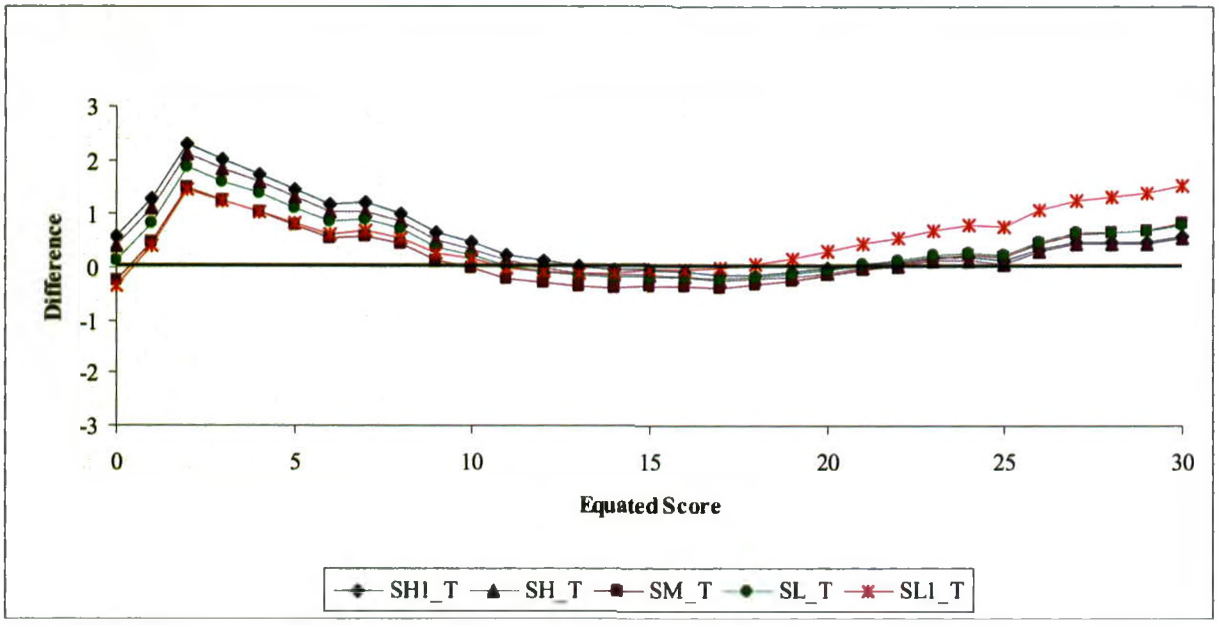
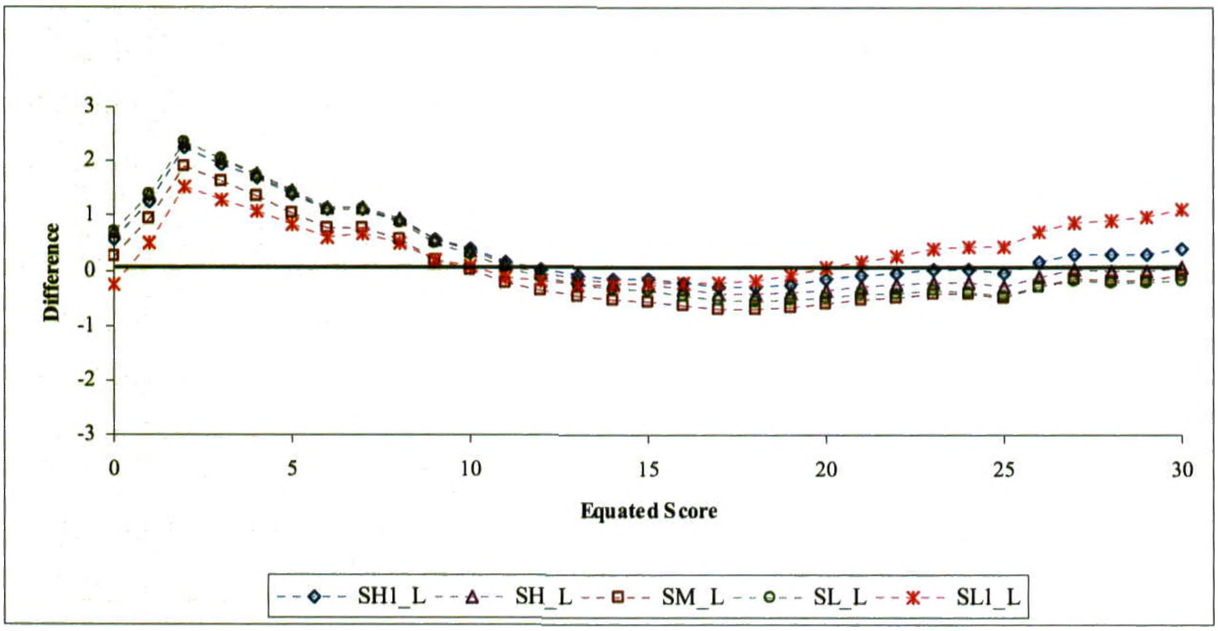


FIGURE 6. Differences of unrounded equated scores between various equatings and criterion equating: Form Similar & Ability Similar – Levine



Figures 7 and 8 show that the Levine observed-score method generally outperforms the Tucker method when the new and old forms are similar but the groups are dissimilar. Furthermore, when the test forms are dissimilar but the groups are similar, both equating methods produce similar results except for the SL1 equating condition in which the Levine observed-score method yields larger differences at the high end of the scale (see Figures 9 and 10). Moreover, when both the new and old forms and groups are dissimilar, both equating methods yield large differences across the scale (see Figures 11 and 12). The Tucker method consistently leads to high equated scores when the new group has a lower ability than the old group. The situation could reverse when a higher ability group is equated to a lower ability group. In addition, the larger the ability differences between the new and old groups, the larger the equated score differences between the samples and population.

FIGURE 7. Differences of unrounded equated scores between various equatings and criterion equating: Form Similar & Ability Dissimilar - Tucker

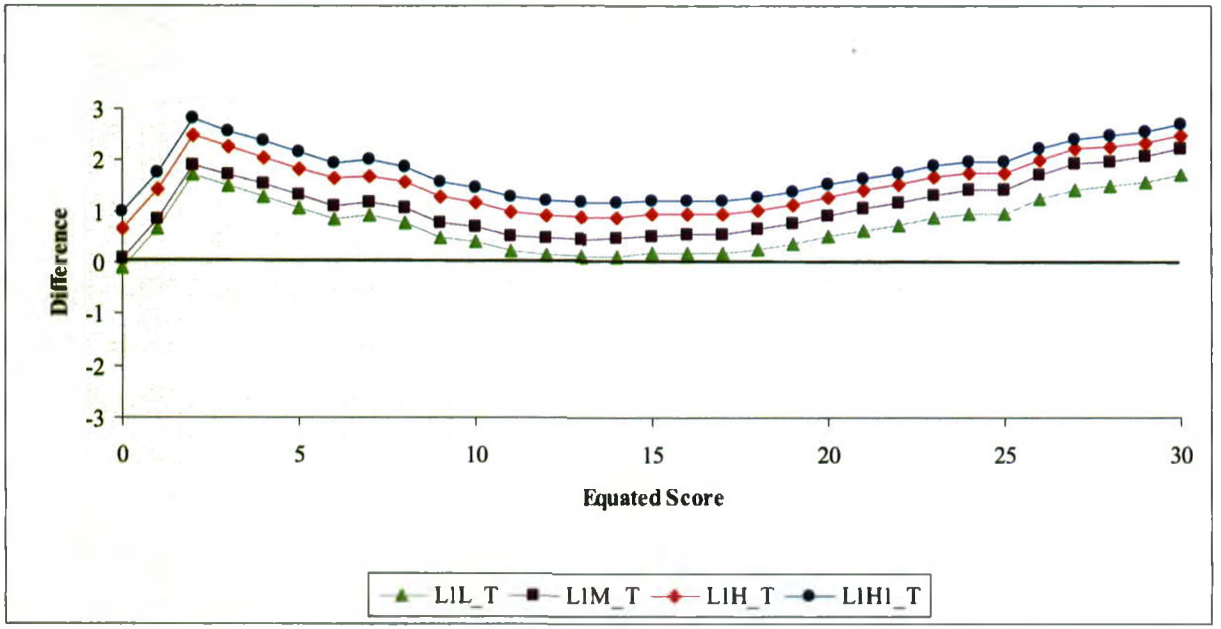


FIGURE 8. Differences of unrounded equated scores between various equating and criterion equating: Form Similar & Ability Dissimilar - Levine

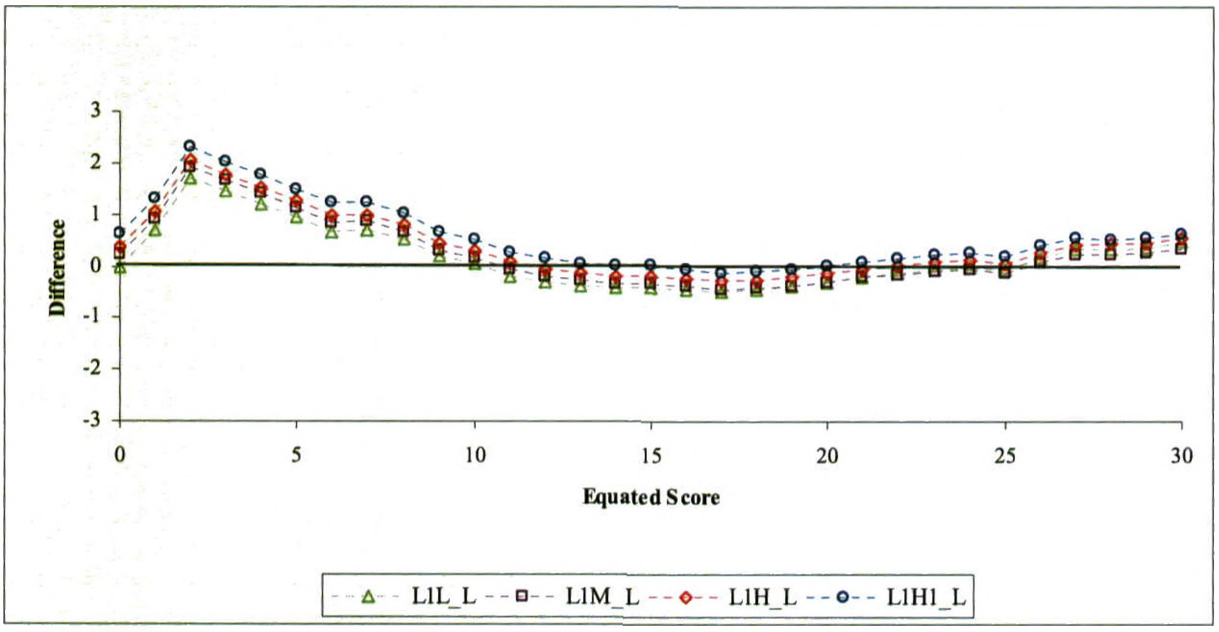


FIGURE 9. Differences of unrounded equated scores between various equatings and criterion equating: Form Dissimilar & Ability Similar – Tucker

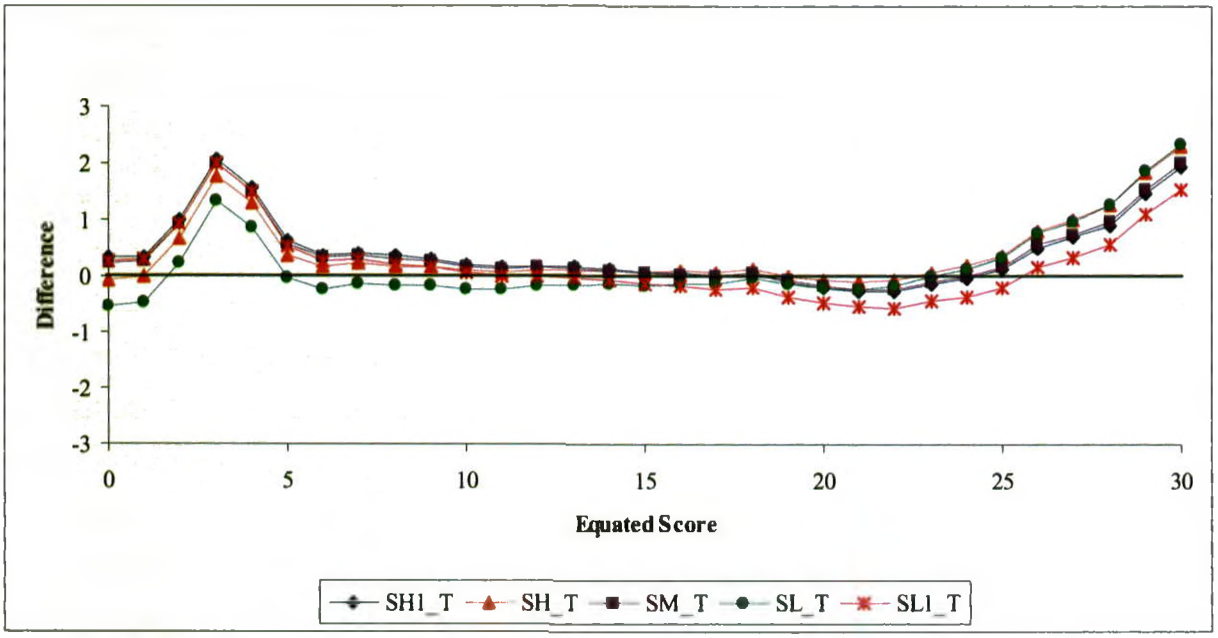


FIGURE 10. Differences of unrounded equated scores between various equating and criterion equating: Form Dissimilar & Ability Similar – Levine

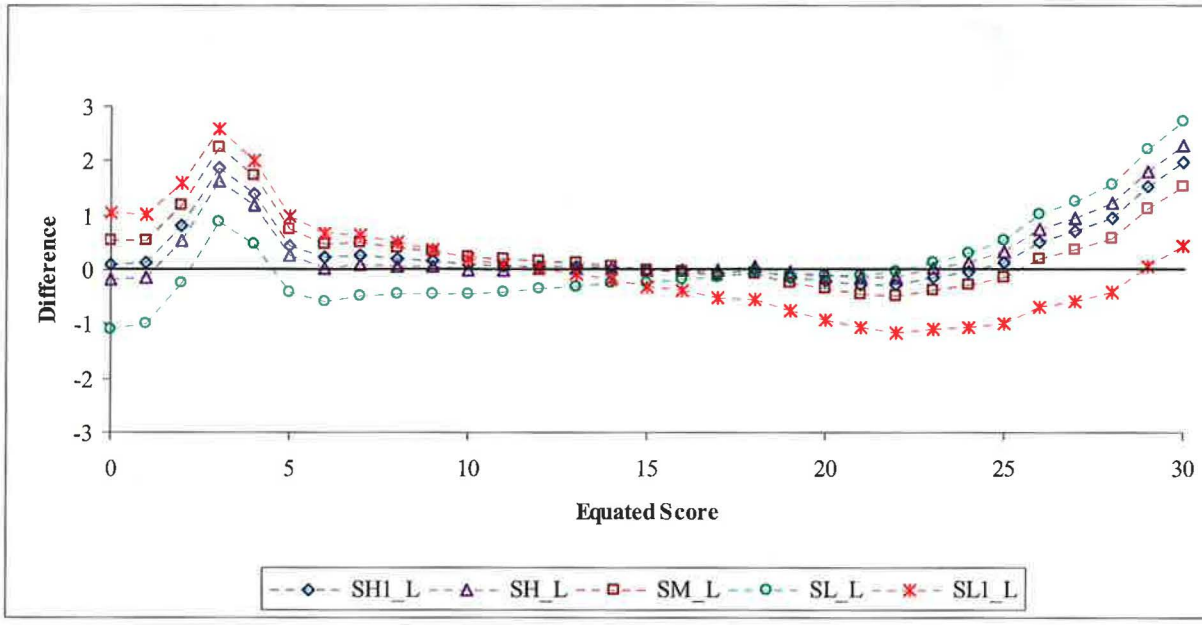


FIGURE 11. Differences of unrounded equated scores between various equatings and criterion equating: Form Dissimilar & Ability Dissimilar – Tucker

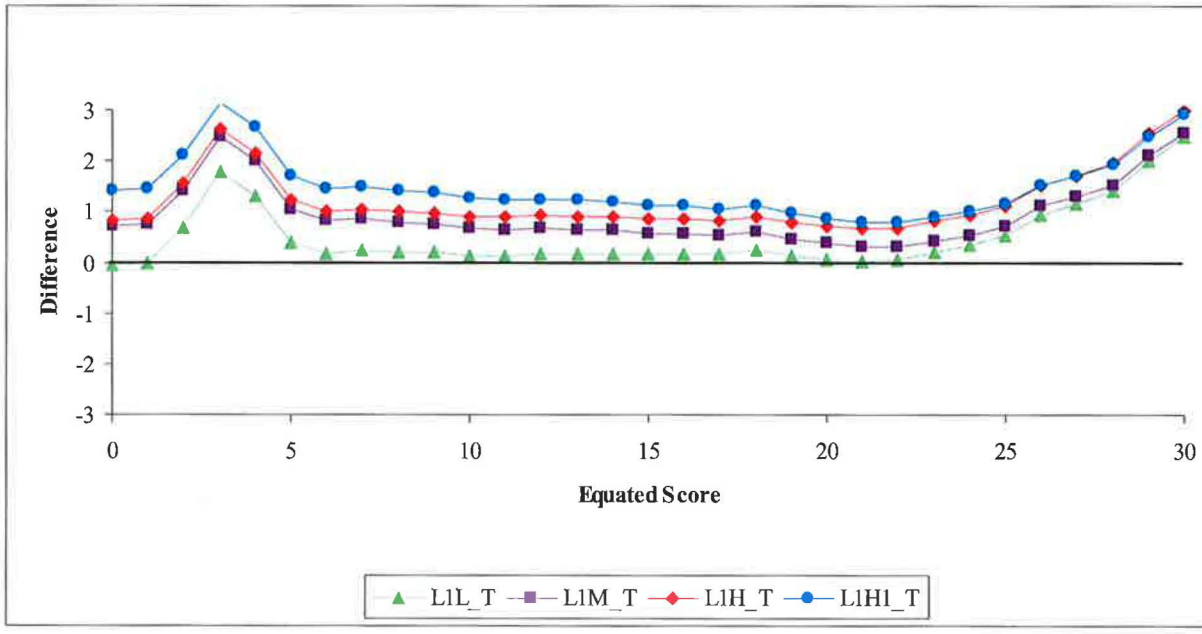
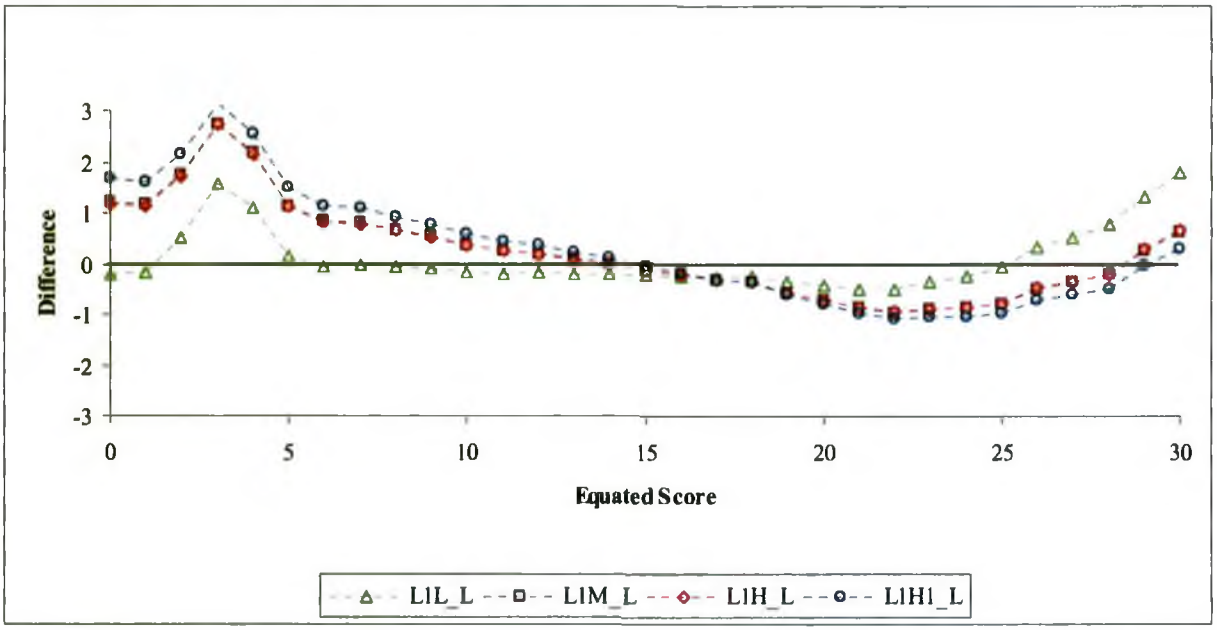


FIGURE 12. Differences of unrounded equated scores between various equatings and criterion equating: Form Dissimilar & Ability Dissimilar – Levine



The findings are in line with the literature about the Tucker and the Levine observed-score equating, indicating when forms are similar in difficulty but groups differ in ability, the Levine observed-score method is preferred to the Tucker method; when forms differ but groups are similar, the Tucker method performs better than the Levine observed-score method. However, the two methods perform similarly when both form and group differences are small. Nevertheless, they should be treated cautiously when both form and group differences are large, as not equating may be the best option.

Unweighted Absolute and Weighted Root Mean Square

The average unweighted absolute equating differences (UAED) and the average weighted root mean square difference (RMSD) between the equated scores from the various equatings and the criterion equating are usually larger under the Tucker method than the Levine observed-score method (see Tables 10 through 17). For example, comparing the UAED for Low1 equated to

High1, High, Mid, Low, and Low1 for forms of similar difficulty, the Tucker method yields 1.797, 1.534, 1.110, 0.762, and 0.596 respectively, but the Levine observed-score method yields only 0.604, 0.511, 0.488, 0.452, and 0.500 to the corresponding equating (see Tables 10 and 11). The similar result can be found for the RMSD. Both the UAED and the RMSD from the Tucker equating are usually larger than those from the Levine observed-score equating. Moreover, comparing for the equatings where the less able new group of Low1 is equated to the more able old groups (i.e., L1L, L1M, L1H, and L1H1), the Levine observed-score method results in consistent UAEDs or RMSDs as the new and old group differences increase. That is, the change of the old group abilities does not affect the Levine observed-score equating. When the Tucker method is used, however, both the UAEDs and the RMSDs become larger as the group differences between the new and old groups increase. In addition, these tables also show that when the new and old group abilities are the same (e.g., High1 to High1) and the test forms are similar or dissimilar in difficulty, both equating methods yield similar UAEDs or RMSDs.

TABLE 10

Average Unweighted Absolute Differences for Similar Forms –Tucker

Tucker	**Similar Forms (Form A to Form B)**				
New	High1	High	Mid	Low	Low1
High1	**0.586**				
High	0.763	**0.526**			
Mid	0.980	0.681	**0.455**		
Low	1.500	1.218	0.789	**0.508**	
Low1	1.797	1.534	1.110	0.762	**0.596**

TABLE 11

Average Unweighted Absolute Differences for Similar Forms –Levine

Levine	Similar Forms (Form A to Form B)				
New	Highl	High	Mid	Low	Lowl
Highl	**0.535**				
High	0.625	**0.584**			
Mid	0.561	0.535	**0.585**		
Low	0.705	0.683	0.727	**0.658**	
Lowl	0.604	0.511	0.488	0.452	**0.500**

TABLE 12

Average Unweighted Absolute Differences for Dissimilar Forms –Tucker

Tucker	Dissimilar Forms (Form C to Form D)				
New	Highl	High	Mid	Low	Lowl
Highl	**0.484**				
High	0.717	**0.445**			
Mid	0.972	0.691	**0.464**		
Low	1.194	0.924	0.639	**0.451**	
Lowl	1.482	1.223	0.932	0.514	**0.450**

TABLE 13

Average Unweighted Absolute Differences for Dissimilar Forms –Levine

Levine	Dissimilar Forms (Form C to Form D)				
New	Highl	High	Mid	Low	Lowl
Highl	**0.420**				
High	0.486	**0.401**			
Mid	0.642	0.509	**0.516**		
Low	0.604	0.471	0.479	**0.597**	
Lowl	0.900	0.725	0.735	0.422	**0.740**

TABLE 14

Average Weighted Root Mean Square Differences for Similar Forms –Tucker

Tucker	Similar Forms (Form A to Form B)				
New	Highl	High	Mid	Low	Lowl
Highl	**0.061**				
High	0.087	**0.056**			
Mid	0.134	0.085	**0.066**		
Low	0.233	0.180	0.106	**0.058**	
Lowl	0.315	0.266	0.192	0.122	**0.095**

TABLE 15

Average Weighted Root Mean Square Differences for Similar Forms –Levine

Levine	Similar Forms (Form A to Form B)				
New	Highl	High	Mid	Low	Lowl
Highl	**0.059**				
High	0.072	**0.074**			
Mid	0.073	0.085	**0.110**		
Low	0.078	0.079	0.095	**0.092**	
Lowl	0.063	0.056	0.068	0.073	**0.067**

TABLE 16

Average Weighted Root Mean Square Differences for Dissimilar Forms –Tucker

Tucker	Dissimilar Forms (Form C to Form D)				
New	Highl	High	Mid	Low	Lowl
Highl	**0.047**				
High	0.091	**0.046**			
Mid	0.142	0.087	**0.044**		
Low	0.190	0.138	0.078	**0.052**	
Lowl	0.249	0.198	0.135	0.059	**0.075**

TABLE 17

Average Weighted Root Mean Square Differences for Dissimilar Forms –Levine

Levine	Dissimilar Forms (Form C to Form D)				
New	High1	High	Mid	Low	Low1
High1	**0.041**				
High	0.047	**0.041**			
Mid	0.079	0.054	**0.056**		
Low	0.078	0.057	0.059	**0.037**	
Low1	0.128	0.106	0.107	0.064	**0.128**

Discussion

This study evaluated the Tucker and the Levine observed-score equating methods when both the new and old examinee group abilities and the new and old form difficulties vary. Three major findings are summarized below.

First, when forms and groups are both similar, research (Kolen, 1990; von Davier, 2008) suggests that both equating methods can produce the same results. The present study supports this conclusion. Although the raw score differences are decomposed differently, the average unrounded scale score differences are not very different between the equating methods.

Second, when forms are similar in difficulty but the groups differ in ability, Kolen and Brennan (2004) recommended that the Levine observed-score method is more appropriate than the Tucker method. The results of the present study also indicate that the Levine observed-score method produces more stable equating results than the Tucker method. When the raw score difference is decomposed into form and group differences, form differences are consistent across group differences for the Levine observed-score method, but increase for the Tucker method as group differences increase. That is, with no change in form difference, the Tucker method overestimates the form differences as group differences increase.

Third, when forms are dissimilar in difficulty and groups vary in ability, the present study suggests that the Levine observed-score method produces more accurate equating results than the Tucker method. Consistent with the form difference decompositions, both the average unweighted equating difference (UAED) and the average weighted root mean square difference (RMSD) are found to be smaller for the Levine observed-score method than the Tucker method. Thus, this study suggests that the Levine observed-score method is more robust than the Tucker method.

A possible reason for this finding is that, under both similar and dissimilar form difficulty conditions, the regression assumptions on which the Tucker method is based on might be violated. The Tucker method assumes that the regression of total raw scores on common-item scores in each form should be the same for both groups. In practice, it is impossible to examine the assumption because each group only takes one form. The present study used simulation techniques to examine the regression assumption for the equating condition when the new and old groups are very different (i.e., L1H1). However, the results of the present study did not show the violation of the regression assumption. The slopes and intercepts of the regressions are very close for the equating conditions of L1H1. Figure 13 and Table 18 show the relationships, slopes, and intercepts for the regressions. Future research is needed to investigate why the Levine observed-score method outperforms the Tucker method.

FIGURE 13. Common Item Regression to Total Raw Score for Similar Forms and Dissimilar Groups (L1H1).

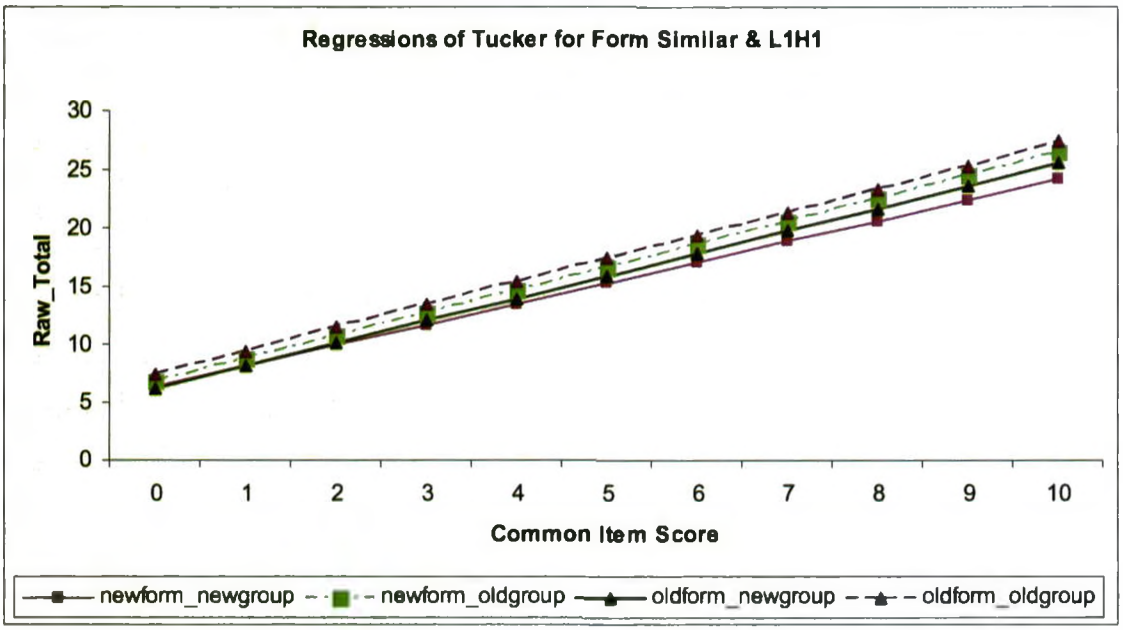


TABLE 18

Regressions for Tucker Equating: Similar Forms and Different Groups

		Intercept	Slope
New Form	New Group (L1)	6.35	1.79
	Old Group (H1)	6.75	1.98
Old Form	New Group (L1)	6.28	1.93
	Old Group (H1)	7.46	1.99

The findings of this study shed light on which equating method should be selected in practice. Some researchers take the position that the Levine observed-score method should be adopted under sizable group difference condition whereas the Tucker method should be adopted under sizable form difference condition. Under the form difference conditions of this study, the results suggest that the Levine observed-score method performs consistently across the conditions of varying group ability differences. That is, the Levine observed-score method

decomposes the form difference consistently under both form difficulty conditions as the group differences change. In contrast, the application of the Tucker method seems only appropriate when group difference is small (for example, less than 0.25 SD), and it may produce inaccurate equating results when group difference becomes larger (for example, 0.5 SD or higher).

It is not clear whether the smaller expansion factor (or γ_2) for the Tucker method might have caused the underestimation for the group differences. However, the Levine observed-score method produces less biased equating results in terms of equated score means, absolute equated differences, and RMSD, which may be related to its larger expansion factor. The performance difference between the equating methods seems directly related to how group and form differences are decomposed, which can be further investigated in future research.

In summary, this study compared the Tucker and the Levine observed-score equating methods under two form difficulty conditions with five group ability conditions. The findings of this study suggest that when group differences in ability are large, the Levine observed-score method is more accurate than the Tucker method in both the estimated decomposed form differences and the equating results. Future research can be undertaken to see if the findings of this study hold for different conditions (e.g., different sample sizes) and tests.

References

Dorans, N. J. & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement,* 37(4), 281-306.

Han, K. T. (2007). *WinGen: Windows Software That Generates Item Response Theory Parameters and Item Responses.* Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment. Retrieved from http://www.umass.edu/remp/software/wingen/.

Lord, F. (1980) Applications of item response theory to practical testing problems. Hillsdale, NJ:Lawrence Erlbaum Associates

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*(1), 97-104.

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

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement,* 7, 201-210.

von Davier, A. (2008). New Results on the Linear Equating Methods for the Non-Equivalent-Groups Design. *Journal of Educational and Behavioral Statistics,* 33(2), 186-203.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2006). *A Comparison of the Frequency Estimation and Chained Equipercentile Methods Under the Common-Item Non-Equivalent Groups Design.* CASMA Research Report, 17. Iowa City, IA: The University of Iowa.

Zimowski, M. F., Muraki, E., Mislevy, R. Jo., & Bock, R. D. (1996). *BILOG-MG: Multi-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software.