Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 28

The Effect of Different Factors on Group Invariance in a Concordance Context with a Single Group Design*

Nooree Huh
Won-Chan Lee†

February 2009

*The authors are grateful to Robert L. Brennan for his comments and help with modifying LEGS for our research purpose.

†Nooree Huh is Research Associate, ACT, Inc., 500 ACT Drive, P.O. Box 168, Iowa City, IA 52243 (email: nooree.huh@act.org). Won-Chan Lee is Associate Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu).

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

Contents

List of Tables

List of Figures

Abstract

Group invariance is one of the important properties of linking scores on different tests. In a concordance situation, in which scores on tests that differ somewhat in content, difficulty, and/or test length are linked, group invariance is likely to be violated. This study examined group invariance under several concordance conditions constructed by a simulation method. The linear, parallel linear, unsmoothed equipercentile, and postsmoothed equipercentile methods were used to obtain conversions. Results of this study indicated that when two tests were dissimilar in terms of test difficulty and test length, the resulting conversions were more group dependent. The results based on a large number of items with a large sample size tended to be less group dependent than those based on a small number of items with a small sample size. In addition, linking scores on an easy test to those on a hard test (for equipercentile linking) and linking scores on a long test to those on a short test were less group dependent than those involving opposite directions.

1 Introduction

Types of test linking have been defined by several researchers (Mislevy, 1992; Linn, 1993; Hanson, Harris, Pommerich, Sconing, & Yi, 2001; Dorans, 2004; Kolen & Brennan, 2004). Kolen and Brennan (2004) summarized Mislevy and Linn's categories based on degrees of similarity between the tests to be linked: equating, vertical scaling, concordance, projection, and statistical moderation. Equating refers to a situation in which scores on tests to be linked are the same in terms of the inferences to be drawn, constructs to be measured, populations to be used, and measurement conditions. By contrast, concordance could differ in measurement conditions (or characteristics) including test length, test format, and administration conditions. Vertical scaling and concordance are somewhat similar in terms of the inferences to be made and construct to be measured. However, unlike vertical scaling, which is based on tests that are designed to be used for dissimilar populations, concordance generally applies to tests that are designed for the same or a similar population.

Group invariance is one of the important properties of linking scores on different forms of a test and can be used to characterize linking relationships. Group invariance can be defined as the extent of the dependence of linking functions on different subgroups. Dorans and Holland (2000) mentioned that group invariance is the most critical requirement for test equating, and proposed that group invariance be a criterion to evaluate whether or not scores that are supposed to be used interchangeably are in fact interchangeable.

When linking results are judged to be group dependent, the conversions based on different subgroups will be different. Using separate conversions for different subgroups might be reasonable when group dependency exists because different conversions may represent different subgroups better. However, in practice, developing separate linking functions for each possible subgroup will be almost impossible due to political issues and the intractably large number of possible subgroups, at least in theory. If a different conversion is adopted for different subgroups, examinees with the same test score will receive different reported scores depending upon their subgroup memberships. Even though assigning different scores to examinees from different subgroups might be argued to be a "fair" treatment, it will raise much more misunderstanding and fairness issues among test takers or stakeholders. In practice, it seems almost always inevitable to use a single conversion for all examinees regardless of their subgroup memberships. From this fairness perspective, the group invariance property for linking is important.

Unlike equating, concordance is more likely to be group dependent. Even though the group invariance property is likely to be violated, concordance is frequently conducted in practice. The most well-known concordance situation may be the linkage between scores of ACT and SAT (Dorans, Lyu, Pommerich, & Houston, 1997). A concordance table between ACT and SAT could provide assistance in an admission or selection decision and also give students and schools a flexible choice of which test to take or administer.

To minimize group dependency of linking, several factors that might affect

the extent of group invariance need to be considered before performing concordance. According to previous group invariance studies in a concordance context (Huh & Kolen, 2006; Yin, Brennan, & Kolen, 2004; Dorans & Holland, 2000; Dorans, Holland, Thayer, & Tateneni, 2003; Tateneni & Dorans, 2003; Pommerich, Hanson, Harris, & Sconing, 2004), such factors as test length, test content, and sample size seem to influence the extent of group dependence of linking.

The study by Huh and Kolen (2006) examined the effect of test length (the number of items), test content, and level of disattenuated correlations on the extent of group invariance in linking scores on tests in a concordance context. To examine the extent of group invariance, Huh and Kolen (2006) created eight concordance contexts based on test length similarity and test content similarity. Their study showed that the number of items affects the extent of group invariance-i.e., group invariance is more likely to be satisfied when linking tests are similar in length in a concordance context. In addition, the results of their study revealed that group invariance was influenced by the level of the disattenuated correlation.

One limitation of the study by Huh and Kolen (2006) and many other group invariance studies in a concordance context is that they were based on real data and thus various levels of conditions such as test length, disattenuated correlations, and test difficulty could not be considered. One of the advantages of simulation is that it can manipulate more conditions than is possible in a real situation and it also can separate the effect of one factor from another to some extent.

The primary purpose of the present paper is to extend Huh and Kolen's (2006) study by applying a simulation method to examine the degree of group invariance for the four linking methods (which will be discussed in the next section) in a concordance context by manipulating factors such as test length similarity, level of test difficulty, sample size, and linking direction (e.g., short to long vs. long to short and easy to hard vs. hard to easy). Previous studies (e.g., Dorans, Holland, Thayer, & Tateneni, 2003; Huh & Kolen, 2004) examined group invariance using various subgroups that were determined based on the racial/ethnical background, gender, geographic region, etc. Dorans (2004) pointed out that group dependency occurs due to the interaction between the relative difficulty of the two tests to be linked and the group membership. One possible factor that could cause the interaction might be the difference in ability among subgroups. Yi, Harris, and Gao (2007) found that if the ability level of subgroups was related to the performance on the test, then equating results were more group sensitive. In this study, group invariance is evaluated based on the performance of two subgroups that differ in the level of ability.

2 Method

The four linking methods employed in this paper are briefly discussed in this section. The simulation procedures are described, followed by the criteria used

to evaluate the degree of group invariance.

2.1 Linking Methods

The four linking methods employed in this paper include the linear, parallel linear, unsmoothed equipercentile, and postsmoothed equipercentile methods. When scores on two tests are linearly related, the linear method is appropriate to be used. Scores that are equal in distance from their means in standard deviation units are set equal in a linear method (Kolen & Brennan, 2004). The linear conversion equation for subgroup k is given by (Kolen & Brennan, 2004):

$$l_{Yk}(x) = \mu_k(Y) + \frac{\sigma_k(Y)}{\sigma_k(X)} \left[x - \mu_k(X)\right], \tag{1}$$

where $l_{Yk}(x)$ is the linear conversion equation for transforming observed score x on Test X to the scale of Test Y for subgroup k; $\sigma_k(X)$ and $\sigma_k(Y)$ are the standard deviations of Test X and Test Y, respectively, for subgroup k; and $\mu_k(X)$ and $\mu_k(Y)$ are mean values of Test X and Test Y, respectively, for subgroup k.

A variation of the linear method is the parallel linear method, which was developed by Dorans and Holland (2000) for the purpose of analytic simplicity. Unlike the linear method, the parallel linear method groups all differences in conversions for subgroups into intercept differences. As Yi, Brennan and Kolen (2004) indicated, using the parallel linear method probably understates subgroup differences with regard to the total group if the group invariance property is not satisfied since slope differences will not be reflected in the conversion. The parallel linear conversion equation for subgroup k is as follows:

$$pl_{Yk}(x) = \mu_k(Y) + \frac{\sigma(Y)}{\sigma(X)} \left[x - \mu_k(X)\right], \tag{2}$$

where $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of Test X and Test Y, respectively, for the total group; and the other terms are the same as defined in Equation 1.

In the equipercentile method, when scores on Form X are linked to those on Form Y, the distribution of scores on Form X transformed to the Form Y scale is set equal to the distribution of scores on Form Y in the population (Kolen & Brennan, 2004). Unlike the linear or parallel linear methods, the equipercentile method does not assume linearity in differences in difficulty among tests. Since a sample data is used to estimate equivalents, it is likely that estimation is not precise. If the score distribution and equipercentile relationships based on the sample are plotted, they will appear irregular, especially when sample size is small. If a "sufficiently large" sample size or the entire population is used, the score distribution and equipercentile relationships likely will be smooth. Statistical smoothing methods are designed to reduce such sampling error.

Before explaining smoothing methods, the term "error" needs to be addressed. There is random error, which occurs due to sampling of examinees to

estimate the population value. By contrast, systematic error can occur in several situations. Kolen and Brennan (2004) list some cases when systematic error occurs: when an estimation method introduces bias in estimating the linking relationship, when statistical assumptions are violated, when a data collection design is improperly implemented, and a group of examinees used to obtain equivalents are different from those who take an equated form. Total error refers to the combination of both random and systematic errors.

Smoothing can reduce random error. In postsmoothing, equipercentile equivalents are smoothed directly. In particular, the cubic spline smoothing method (with a degree of 1.0) is employed in this paper (Kolen & Brennan, 2004). As Kolen and Brennan (2004) indicate, smoothing can lead to less random error in estimating equipercentile equivalents than the unsmoothed equipercentile method. However, it is also possible that smoothing produces more total error than the unsmoothed equipercentile method due to more systematic error.

2.2 Simulation Procedure

The data collection designs used for equating can also be used for concordance. Many of the concordance studies employed a single group design (Dorans & Holland, 2000; Yin et al., 2004; Huh & Kolen, 2006). Thus, in the present study, data are simulated under a single group design.

In order to create simulation conditions that are as realistic as possible, item parameters from real test data were initially used. Data were from multiple forms of a large-scale mathematics test. All calibrated items from these forms constituted an item pool, which were used in the simulation to create tests to be linked. The created tests were different in terms of length and/or difficulty.[1] In addition, the tests to be linked were constructed to have no overlapping items. Item parameters were estimated using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) assuming a three-parameter logistic (3PL) IRT model. The estimated item parameters were treated as "true" parameters that were used to simulate item responses. The simulation and analyses involved the following steps.

First, two tests (Test X and Test Y) were created for each of 14 concordance pairs with varying levels of test length (number of items) and test difficulty (easy, moderate, and hard) based on the "true" b parameters. Table 1 summarizes these 14 concordance pairs, in which Test X is linked to Test Y. All items available for simulation were sorted in an increasing order by the "true" b parameters. Based on the sorted items, for example, the first 120 items were used to create a pair of easy 60-item tests for Test X and Test Y. Note that before selecting items, the first few items in the sorted list were excluded such that a mean b-parameter value for an easy-item pool could be set to approximately zero. Two hard tests were constructed in a similar manner using the last 120 items from the sorted list, and the target mean b-parameter value for each test

[1] In this regard, the relationship to link two tests considered in the simulation study is concordance rather than equating (Dorans, 2004).

Table 1: Linking Pairs

Condition	Test X	Test Y
C1	Easy 20 items	Hard 20 items
C2	Hard 20 items	Easy 20 items
C3	Easy 60 items	Hard 60 items
C4	Hard 60 items	Easy 60 items
C5	Moderate 20 items	Moderate 60 items
C6	Moderate 60 items	Moderate 20 items
*C7–C10	Moderate 20 items	Moderate 20 items
*C11–C14	Moderate 60 items	Moderate 60 items

* Sample size varies across each condition.

was 1.0. Items with moderate difficulty were selected in a random manner to include both hard and easy items. With the target mean b-parameter value of 0.5, moderate-difficulty tests were created to reflect a real situation where tests to be linked have both hard and easy items. Note that the mean of the b parameters in the pool was approximately 0.2. Twenty-item tests were selected from these 60-item tests. The mean and standard deviation of b-parameters for the tests created for simulation are provided in Table 2.

Second, the same sample of examinees was assumed to take both tests under a single group design. Within a group of examinees, two subgroups were defined based on the examinees' ability levels (θ). These two ability subgroups were evaluated for group invariance of concordance. Three levels of subgroup difference were considered by drawing θ values for the two subgroups from two different distributions. Table 3 shows the θ distributions used to draw samples for the two subgroups for the three levels of subgroup difference. In this study, four levels of sample size were considered for C7 through C10 and also for C11 through C14 listed in Table 1: 1) 200 for each subgroup per test; 2) 1000 for each subgroup per test; 3) 200 for a high ability group and 1000 for a low ability group per test; and 4) 1000 for a high ability group and 200 for a low ability group per test. For other concordance pairs, a sample size of 1000 for each subgroup was used. One thousand samples of θ were drawn from each of the six distributions listed in Table 3. For the conditions of a smaller sample size, the first 200 θ values from the 1000 samples were used. Table 4 provides the mean and standard deviation of θ values for the samples drawn from each of the six distributions.

Third, the response vectors of examinees for each sample were obtained for each concordance condition using θ and item parameters obtained from the above steps. When the 3PL probability computed using the "true" item parameters and θ values was larger than a random number drawn from a uniform distribution, the response for that item was recorded as 1, otherwise 0.

Fourth, the four linking methods were conducted for each of the 42 linking

Table 2: Mean and Standard Deviation of b-parameters for the Tests Used for Simulation

Test	Mean	SD	Min	Max
Test X, Easy 60	-0.036	0.857	-2.943	2.111
Test Y, Easy 60	0.021	0.820	-1.792	2.243
Test X, Easy 20	-0.103	1.003	-2.943	1.988
Test Y, Easy 20	-0.046	0.812	-1.792	1.987
Test X, Moderate 60	0.551	1.399	-2.352	2.985
Test Y, Moderate 60	0.516	1.394	-2.382	2.775
Test X, Moderate 20	0.444	1.410	-2.352	2.644
Test Y, Moderate 20	0.412	1.414	-2.382	2.579
Test X, Hard 60	1.121	0.742	-0.004	2.775
Test Y, Hard 60	1.149	0.760	0.016	2.985
Test X, Hard 20	1.077	0.725	-0.004	2.579
Test Y, Hard 20	1.097	0.730	0.016	2.644

Table 3: Distributions of θ for Three Levels of Subgroup Ability Differences

	Low Ability Subgroup	High Ability Subgroup
No Difference	N(0,1)	N(0,1)
Small Difference	N(-0.05,1)	N(0.05,1)
Large Difference	N(-0.3,1)	N(0.3,1)

pairs (i.e., 14 concordance pairs listed in Table 1 times the three levels of group difference). All analyses were conducted using the computer program LEGS (linking with equivalent groups or single group design, Brennan, 2004).

Fifth, the above steps were repeated 100 times. The average *REMSD* and average $RMSD(x)$ were computed for each linking pair (discussed in the next section). To evaluate the relative magnitude of summary statistics, "Difference that Matters" (DTM), which was introduced by Dorans, Holland, Thayer, and Tateneni (2003), was computed.

Even though the statistical methodologies used for concordance and equating are the same, what distinguishes between concordance and equating are the similarity of the content area to be measured, item types, the number of items, and/or the difference in difficulty between tests to be linked. In this paper, the tests to be linked are created to be different in terms of the number of items and difficulty level. Hence, the linking conditions are more likely to be considered as concordance rather than equating. In addition, since a single group of examinees take both tests, it is assumed that the two tests are built to be used for the same

Table 4: Mean and Standard Deviation of θ Values for the Samples Used for Simulation

Sample Size	Difference in Ability Between Subgroups	Low Ability Subgroup	High Ability Subgroup
1000	None	0.000 (0.992)	0.001 (1.060)
	Small	-0.080 (1.011)	0.061 (0.990)
	Large	-0.267 (1.024)	0.268 (0.993)
200	None	-0.023 (1.030)	-0.047 (1.068)
	Small	-0.116 (1.032)	0.036 (0.897)
	Large	-0.277 (1.059)	0.282 (1.011)

Note. Numbers in parentheses represent standard deviation.

population. In this regard, each linking can be considered as concordance rather than vertical scaling.

In summary, this study examines the effects of the test length, test difficulty, sample size, subgroup ability level difference, and direction of linking on the extent of group invariance using a simulation method. In this study, subgroups are divided by the ability level (θ) to investigate group invariance in a concordance context. The summary statistics of *REMSD* and $RMSD(x)$ are used to examine the extent of group invariance as discussed next.

2.3 Criteria

Dorans and Holland (2000) developed statistics to summarize differences between the equivalents obtained from subgroups and the total group. Among them are the standardized Root Mean Square Difference for a specific raw score x, $RMSD(x)$; the standardized Root Expected Mean Square Difference, *REMSD*, which summarizes the overall differences for the entire group; and an equally weighted *REMSD*, *ewREMSD* (Kolen & Brennan, 2004), which uses the same weight for all score points. *REMSD* and $RMSD(x)$ are considered in the present paper. In this study, $t_Y(x)$ represents transformed scores on Test X to the scale of Test Y for a total group, and $t_{Yk}(x)$ represents transformed scores on Test X to the scale of Test Y for subgroup k using any of the four linking methods used in this study. Let N_k be the sample size for subgroup k, N the total number of examinees, and $w_k = N_k/N$ the weight for subgroup k. Then,

$$RMSD(x) = \frac{\sqrt{\sum_{k=1}^{K} w_k [t_{Yk}(x) - t_Y(x)]^2}}{\sigma(Y)}, \tag{3}$$

where $\sigma(Y)$ is the standard deviation of the total group who took Test Y, and K is the number of subgroups ($K = 2$ in this study).

Let $min(x)$ and $max(x)$ be the observed minimum and maximum values of scores on Test X, N_{xk} the number of examinees for subgroup k with a particular score on Test X, and $v_{xk} = N_{xk}/N_k$ a weighting factor for subgroup k on score x. A computational formula for *REMSD* is:

$$REMSD = \frac{\sqrt{\sum_{k=1}^{K} w_k \sum_{min(x)}^{max(x)} v_{xk}[t_{Yk}(x) - t_Y(x)]^2}}{\sigma(Y)}. \quad (4)$$

One hundred replications were performed in this study. To obtain the average $RMSD(x)$ and *REMSD* over 100 replications, $t_Y(x)$, $t_{Yk}(x)$, and $\sigma(Y)$ in Equations 3 and 4 were replaced with $\bar{t}_Y(x) = \sum t_Y(x)/100$, $\bar{t}_{Yk}(x) = \sum t_{Yk}(x)/100$, and $\overline{\sigma}(Y) = \sqrt{\sum \sigma^2(Y)/100}$, where the summations were taken over 100 replications.

The extent of group invariance was assessed using the average *REMSD* and $RMSD(x)$ values. In addition, to evaluate the relative magnitude of *REMSD* for unrounded raw scores, "Difference That Matters" (DTM) was employed. The DTM depends on the reporting scale of a test and is defined as a half of a reported score unit for unrounded scores. The DTM can be standardized so that the standardized DTM can be used as a benchmark for evaluating *REMSD*. As Kolen and Brennan (2004) pointed out, however, the DTM needs to be understood as a convenient benchmark, not a strict rule.

3 Results

Table 5 provides the raw score mean and standard deviation for the combined group in each of the 14 concordance pairs used in this study, and Table 6 displays those for each subgroup. The four linking methods (linear, parallel linear, unsmoothed equipercentile, and postsmoothed equipercentile) were applied to each pair and group invariance statistics were obtained. Table 7 shows correlations between linked tests for combined group. Note that the correlation values vary approximately from .67 to .90. The effects of five factors (test difficulty, test length, sample size, linking direction, and the level of subgroups difference) on group invariance were examined statistically and graphically.

Figure 1 provides conversion graphs and cumulative distributions for Condition 3, Condition 4, and Condition 14 for a group with large difference between subgroups. The conversions and cumulative distributions were almost identical regardless of the level of subgroup difference. Note that sets of conditions, Conditions 1 and 3, which are based on linking scores on an easy test to a hard test, and Conditions 2 and 4, which are based on linking scores on a hard test to an easy test, show similar patterns of conversions regardless of the ability difference between subgroups. Thus, only the results for Conditions 3 and 4 are presented here.

Recall that, under Condition 3, an easy 60-item test was linked to a hard 60-item test with the sample size of 1000 for each subgroup, and under Condition

Table 5: Sample Size, Mean, and Standard Deviation of Linked Tests for Combined Group

		Test X						Test Y					
		No-Diff		S-Diff		L-Diff		No-Diff		S-Diff		L-Diff	
Cond.	Sample Size*	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD
C1	2000 (1000,1000)	11.89	4.15	11.89	4.11	11.89	4.18	7.73	3.77	7.69	3.67	7.75	3.80
C2	2000 (1000,1000)	7.63	4.00	7.59	3.91	7.65	4.04	11.36	4.32	11.35	4.28	11.37	4.36
C3	2000 (1000,1000)	34.99	12.19	34.97	12.06	34.99	12.27	23.06	10.46	22.89	10.13	23.11	10.59
C4	2000 (1000,1000)	22.58	10.63	22.46	10.31	22.65	10.75	34.44	12.27	34.41	12.14	34.43	12.35
C5	2000 (1000,1000)	10.36	3.28	10.33	3.21	10.36	3.31	29.03	9.47	28.93	9.23	29.05	9.57
C6	2000 (1000,1000)	29.39	9.34	29.31	9.10	29.43	9.43	9.99	3.42	9.96	3.36	10.01	3.45
C7	400 (200,200)	10.29	3.31	10.24	3.12	10.39	3.36	9.92	3.46	9.88	3.27	10.03	3.51
C8	1200 (1000,200)	10.35	3.33	10.44	3.21	10.83	3.31	9.99	3.49	10.08	3.36	10.50	3.45
C9	1200 (200,1000)	10.34	3.23	10.19	3.18	9.91	3.27	9.97	3.38	9.82	3.32	9.52	3.40
C10	2000 (1000,1000)	10.36	3.27	10.32	3.22	10.37	3.30	10.00	3.42	9.96	3.36	10.00	3.46
C11	400 (200,200)	29.15	9.47	28.99	8.72	29.50	9.66	28.79	9.57	28.63	8.84	29.11	9.80
C12	1200 (1000,200)	29.38	9.54	29.64	9.12	30.93	9.51	29.02	9.67	29.27	9.24	30.55	9.61
C13	1200 (200,1000)	29.32	9.17	28.84	8.93	27.96	9.24	28.93	9.31	28.47	9.10	27.57	9.36
C14	2000 (1000,1000)	29.39	9.34	29.29	9.11	29.41	9.44	29.03	9.47	28.92	9.22	29.05	9.56

* The numbers in parentheses indicate the numbers of examinees in high and low ability groups, respectively.
Note. No-Diff = no subgroup difference; S-Diff = small subgroup difference; L-Diff = large subgroup difference; Cond. = condition.
Note. Link direction is from Test X to Test Y for all conditions.

Table 6: Mean and Standard Deviation of Linked Tests for Each Subgroup

	No-Diff				S-Diff				L-Diff			
	High		Low		High		Low		High		Low	
Cond.	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
C1	11.8	7.8	11.9	7.7	12.1	7.9	11.6	7.5	12.9	8.5	10.9	7.0
	(4.2)	(3.8)	(4.1)	(3.6)	(4.1)	(3.7)	(4.1)	(3.6)	(4.0)	(3.9)	(4.1)	(3.5)
C2	7.7	11.4	7.6	11.4	7.8	11.6	7.4	11.1	8.5	12.4	6.8	10.4
	(4.1)	(4.4)	(3.9)	(4.3)	(4.0)	(4.3)	(3.8)	(4.3)	(4.2)	(4.1)	(3.7)	(4.4)
C3	34.9	23.2	35.0	23.0	35.7	23.5	34.2	22.3	38.0	25.5	31.9	20.8
	(12.4)	(10.7)	(12.0)	(10.2)	(12.0)	(10.3)	(12.1)	(9.9)	(11.7)	(11.1)	(12.0)	(9.5)
C4	22.7	34.4	22.5	34.5	23.1	35.2	21.8	33.6	25.1	37.5	20.2	31.4
	(10.9)	(12.5)	(10.4)	(12.0)	(10.5)	(12.1)	(10.1)	(12.2)	(11.2)	(11.8)	(9.7)	(12.1)
C5	10.4	29.1	10.4	29.0	10.5	29.5	10.1	28.4	11.1	31.3	9.7	26.8
	(3.4)	(9.7)	(3.2)	(9.2)	(3.2)	(9.3)	(3.2)	(9.2)	(3.3)	(9.5)	(3.2)	(9.1)
C6	29.4	10.0	29.4	10.0	29.9	10.2	28.7	9.8	31.7	10.8	27.2	9.3
	(9.6)	(3.5)	(9.1)	(3.3)	(9.1)	3.4)	(9.0)	(3.4)	(9.4)	(3.4)	(8.9)	(3.3)
C7	10.3	9.9	10.3	10.0	10.4	10.1	10.1	9.7	11.1	10.8	9.8	9.2
	(3.4)	(3.5)	(3.3)	(3.4)	(3.0)	(3.2)	(3.2)	(3.4)	(3.3)	(3.6)	(3.3)	(3.5)
C8	10.4	10.0	10.3	9.9	10.5	10.2	10.1	9.7	11.1	10.8	9.7	9.2
	(3.3)	(3.5)	(3.3)	(3.4)	(3.2)	(3.4)	(3.2)	(3.3)	(3.3)	(3.4)	(3.3)	(3.4)
C9	10.3	9.9	10.4	10.0	10.4	10.0	10.2	9.8	11.2	10.9	9.7	9.2
	(3.4)	(3.5)	(3.2)	(3.4)	(3.0)	(3.2)	(3.2)	(3.4)	(3.4)	(3.5)	(3.2)	(3.4)
C10	10.4	10.0	10.4	10.0	10.5	10.2	10.2	9.8	11.1	10.8	9.7	9.3
	(3.3)	(3.5)	(3.2)	(3.4)	(3.2)	(3.4)	(3.2)	(3.4)	(3.3)	(3.4)	(3.2)	(3.4)
C11	29.1	28.7	29.2	28.9	29.5	29.2	28.5	28.1	32.1	31.1	27.3	26.7
	(9.6)	(9.7)	(9.3)	(9.4)	(8.5)	(8.6)	(9.0)	(9.1)	(9.6)	(9.4)	(9.2)	(9.1)
C12	29.4	29.1	29.2	28.8	29.9	29.5	28.5	28.1	31.6	31.2	27.3	26.8
	(9.6)	(9.7)	(9.3)	(9.4)	(9.1)	(9.3)	(9.0)	(9.1)	(9.5)	(9.6)	(9.1)	(8.8)
C13	29.1	28.7	29.4	29.0	29.5	29.1	28.7	28.3	31.7	31.4	30.7	30.3
	(9.6)	(9.8)	(9.1)	(9.2)	(8.5)	(8.6)	(9.0)	(9.2)	(10.1)	(9.7)	(9.5)	(9.6)
C14	29.4	29.1	29.4	29.0	29.9	29.5	28.7	28.3	31.7	31.3	27.2	26.8
	(9.6)	(9.7)	(9.1)	(9.2)	(9.1)	(9.2)	(9.0)	(9.2)	(9.4)	(9.5)	(9.0)	(9.1)

Note. Cond. = condition; X = Test X; Y = Test Y; High = high ability group; Low = low ability group; No-Diff = no subgroup difference; S-Diff = small subgroup difference; L-Diff = large subgroup difference.
Note. The numbers in parentheses represent standard deviation.

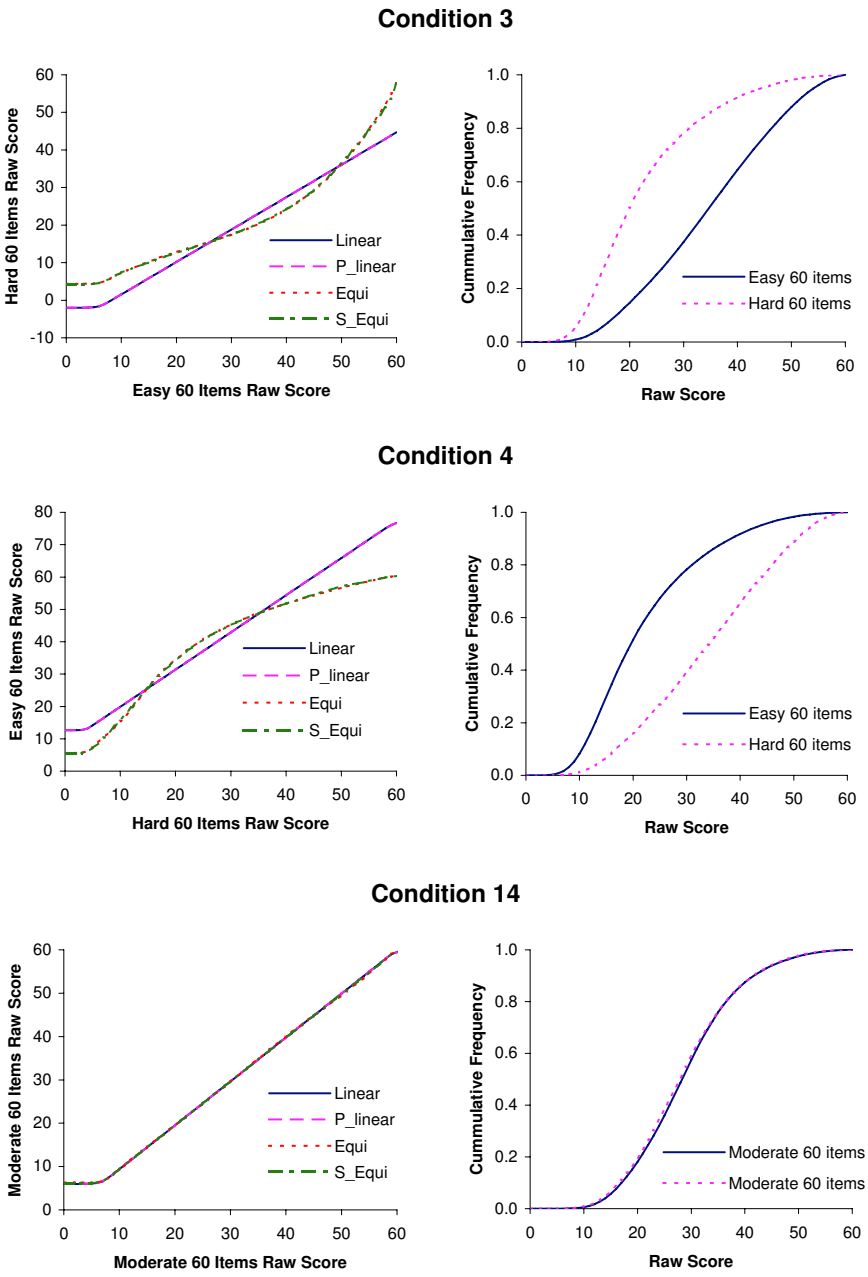

Figure 1: Conversions and Cumulative Frequency Distributions for Large Subgroup Differences (Total Group Based)

Table 7: Correlation between Linked Tests in Each Condition for Total Group

Condition	No-Diff	S-Diff	L-Diff
C1	.733	.723	.735
C2	.754	.746	.756
C3	.858	.855	.859
C4	.868	.866	.869
C5	.785	.774	.788
C6	.799	.790	.803
C7	.712	.674	.723
C8	.716	.696	.717
C9	.696	.685	.699
C10	.707	.695	.712
C11	.891	.871	.896
C12	.892	.882	.892
C13	.883	.877	.883
C14	.887	.880	.889

Note. No-Diff = no subgroup difference;
S-Diff = small subgroup difference;
L-Diff = large subgroup difference

4, an opposite direction linking was performed. Likewise, a set of conditions, Conditions 5 through 14, which are based on linking scores on tests that are similar in the levels of test difficulty, show similar patterns of conversions, and only the results for Condition 14 are reported. Condition 14 involves linking a moderate 60-item test to another moderate 60-item test with the sample size of 1000 for each subgroup. When two tests are different in their test difficulty (i.e., C1 through C4), the cumulative frequencies between the two tests are somewhat different, which results in different conversions among the four linking methods, especially between the linear and nonlinear methods. The two linear methods (i.e., linear and parallel linear) produce almost identical conversions. As can be seen from the cumulative frequency plot, the assumptions of the linear methods do not seem appropriate for the data for C1 through C4. Adjusting only the first two moments (mean and standard deviation) as the linear methods do does not seem to adjust properly the difference in the whole score distributions between the two tests. The two nonlinear methods (unsmoothed and smoothed equipercentile) show very similar conversions to each other. The cumulative frequencies between the two tests with a similar difficulty level (i.e., C5 through C14) are almost identical, which, in turn, results in almost identical conversions among the four linking methods.

The discrepancy in the frequency distributions between two tests to be linked also affects group invariance statistics, which are summarized in Table 8. As can be seen in the table, the *REMSD* values for the linear method for C1 through C4,

which are based on linking scores on tests that are dissimilar in test difficulty, are almost always larger than those for the equipercentile methods. Recall that for C1 through C4, linearity assumptions seemed to be inappropriate. However, the *REMSD* values for the parallel linear method are as good as those for the nonlinear methods, and even better under the small subgroup difference condition. The low values of *REMSD* across all conditions for the parallel linear method under the condition of the small subgroup difference are remarkable. Even if the linear and parallel linear methods produce very similar conversions, the degree of group invariance for the parallel linear method is substantially smaller than that for the linear method. For C5 through C14 where tests are similar in difficulty level, the *REMSD* values for the linear method are similar to those for the other three methods, regardless of the difference in ability between subgroups.

Table 8 also shows that the *REMSD* values, in general, tend to get larger when the difference in ability level between subgroups gets larger. Notice that the magnitude of the *REMSD* values for all four linking methods is largest for C1 through C4 when the subgroup difference is large. However, when the subgroup difference is small, the *REMSD* values sometimes are smaller than they are when the subgroup difference is none. For the condition of the no subgroup difference, the *REMSD* values range from .001 to .014 for the linear method, .000 to .013 for the parallel linear method, .003 to .009 for the unsmoothed equipercentile method, and .002 to .015 for the smoothed equipercentile method.

The *REMSD* values are provided graphically in Figure 2 for a group with large difference in ability between subgroups. When equipercentile methods with and without smoothing are used to link scores on tests with different test difficulty (C1 through C4), the *REMSD* values tend to be smaller for the 60-item tests than those for the 20-item tests. The linking direction also seems to influence the group invariance statistics: the hard-to-easy (C2 and C4) direction results in higher *REMSD* than the easy-to-hard (C1 and C3) direction when the equipercentile methods are used. Note that the linear methods were not considered to examine the effect of the different direction in C1 through C4 since the linearity assumption did not seem to hold for these conditions. In addition, the long-to-short (C6) direction results in smaller *REMSD* than the short-to-long (C5) direction across all four linking methods. The effect of sample size on group invariance can be observed from the bottom of the right side of Figure 2. Conditions C11 through C14 are based on moderate 60-item tests and vary in their sample size. As can be seen from Figure 2, the graph based on C11 through C14 looks somewhat flatter than other five comparison graphs. Figure 2 and Table 8 together suggest that the effect of the sample size is less noticeable than the effect of the test difficulty or test length because the magnitude of changes in the *REMSD* values based on the sample size is less than those based on the test difficulty and test length.

It is noteworthy that reliability seems to affect the degree of group invariance. Reliability of scores for each test in C1 through C6, in which two tests to be linked show differences in reliability, is provided in Table 9. The reliability values were computed using the "true" item and ability parameters based on

Table 8: *REMSD* Values

	No-Diff				S-Diff				L-Diff			
Cond.	L	PL	E	SE	L	PL	E	SE	L	PL	E	SE
C1	.011	.010	.009	.015	.020	.004	.007	.016	.071	.025	.023	.032
C2	.011	.010	.007	.011	.022	.003	.010	.015	.075	.022	.032	.036
C3	.014	.013	.007	.009	.026	.003	.007	.009	.090	.026	.016	.019
C4	.014	.012	.008	.008	.025	.003	.008	.010	.084	.021	.019	.022
C5	.007	.002	.009	.011	.008	.006	.009	.010	.031	.027	.031	.033
C6	.004	.000	.009	.008	.007	.006	.009	.008	.025	.020	.024	.025
C7	.002	.002	.003	.002	.002	.001	.003	.002	.004	.004	.005	.003
C8	.002	.002	.005	.002	.002	.002	.006	.003	.005	.004	.006	.006
C9	.003	.002	.005	.003	.000	.000	.004	.002	.003	.003	.008	.003
C10	.002	.001	.003	.002	.004	.003	.005	.004	.005	.005	.007	.006
C11	.002	.000	.003	.003	.001	.001	.004	.002	.001	.000	.003	.002
C12	.002	.001	.003	.003	.000	.000	.003	.002	.002	.001	.003	.003
C13	.001	.001	.003	.002	.000	.000	.003	.002	.002	.002	.004	.004
C14	.002	.001	.003	.002	.002	.001	.004	.002	.001	.001	.002	.002

Note. L = linear method; PL = parallel linear method; E = unsmoothed equipercentile method; SE = postsmoothed equipercentile method; No-Diff = no subgroup difference; S-Diff = small subgroup difference; L-Diff = large subgroup difference.

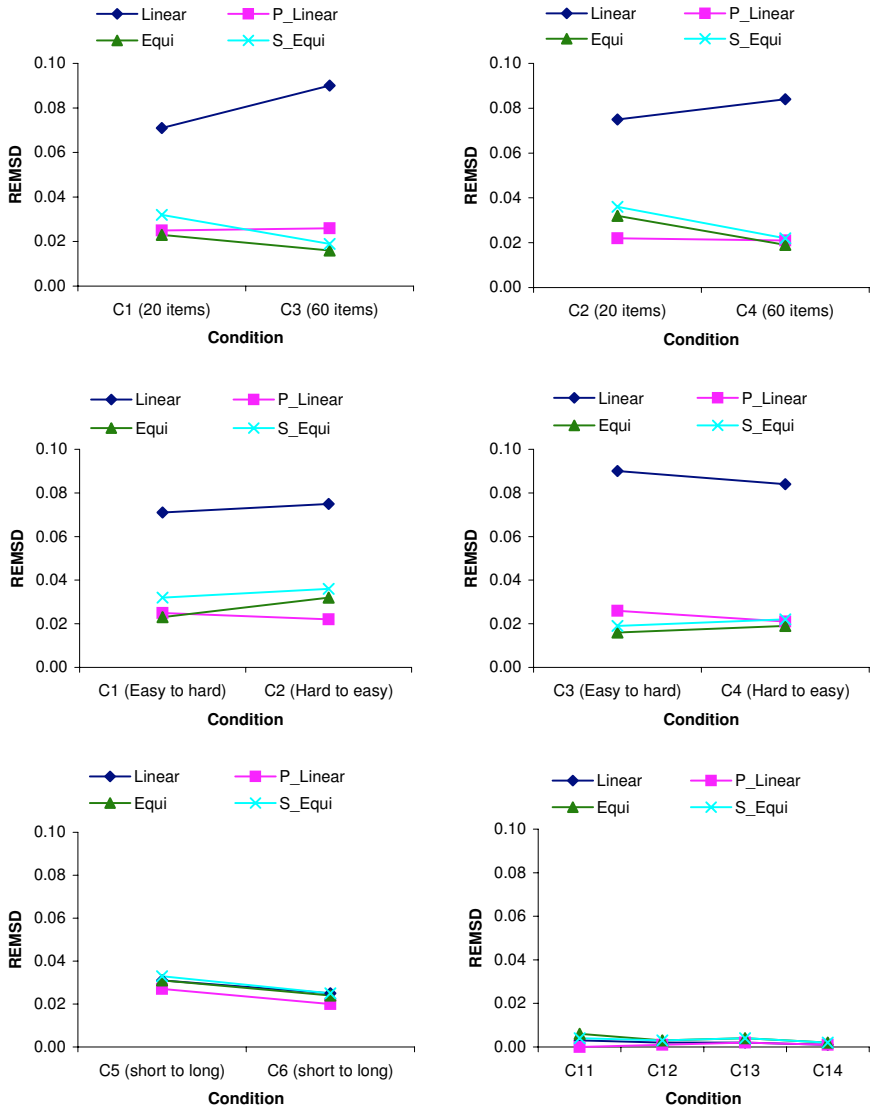

Figure 2: *REMSD* for Large Subgroup Difference

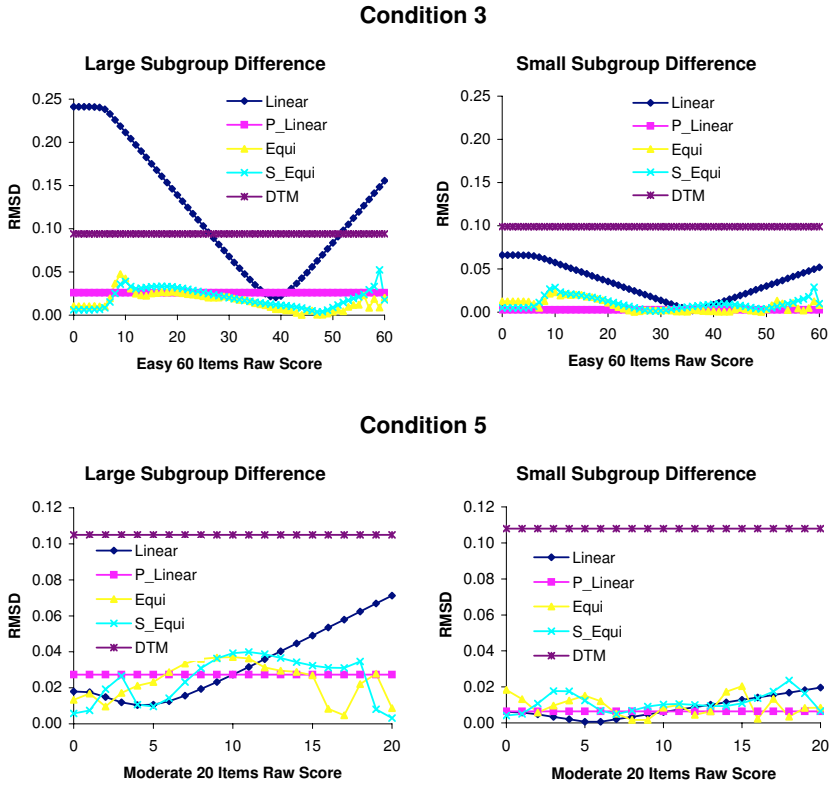

Figure 3: *RMSD* for Conditions 3 and 5 for Total Group

the procedure discussed by Kolen, Zeng, and Hanson (1996).

Recall that the linking direction was from Test X to Test Y. The *REMSD* values from Table 8 and reliability in Table 9 reveal that there is a certain relationship between the degree of group invariance and the direction of linking. When the *REMSD* values are compared between C1 (an easy test linked to a hard test) and C2 (a hard test linked to an easy test) for the equipercentile methods (smoothed and unsmoothed), C1 tends to show smaller *REMSD* values (i.e., more group invariant) than C2 when the subgroup difference exists. This tendency is more evident when C3 (an easy test linked to a hard test) is compared to C4 (a hard test linked to an easy test). C1 and C2 are based on 20-item tests, whereas C3 and C4 are based on 60-item tests. When subgroups have differences in ability, C3 tends to show smaller *REMSD* values than C4 for both equipercentile methods, indicating that the results for C3 are more group invariant than those for C4. As shown in Table 9, C1 and C3 show more group invariant results than C2 and C4, respectively. C1 and C3 represent conditions where a test with relatively higher score reliability is linked to a test with a

Table 9: Reliability of Tests Based on Total Groups

	No-Diff		S-Diff		L-Diff	
Cond.	Test X	Test Y	Test X	Test Y	Test X	Test Y
C1	.802	.731	.797	.715	.805	.736
C2	.771	.813	.758	.809	.775	.816
C3	.930	.898	.928	.891	.931	.901
C4	.902	.931	.895	.929	.904	.932
C5	.691	.889	.678	.883	.697	.892
C6	.885	.722	.879	.711	.888	.727

No-Diff = no subgroup difference; S-Diff = small subgroup difference; L-Diff = large subgroup difference.

relatively lower reliability. The same pattern can be observed for linking a short test to a long test (C5) and a long test to a short test (C6). Note that C6 tends to be more group invariant than C5. A test with a relatively higher reliability is linked to a test with a relatively lower reliability for C6, and the reverse is true for C5.

Unlike *REMSD* which summarizes group invariance across all raw score points, *RMSD* shows group dependency at each score point. *RMSD* was computed and plotted in Figure 3. Because of the limited space and the similarity of results, only the results for two of the fourteen linkings are provided for two levels of subgroup ability difference: large and small difference. By comparing the plots on the left side of Figures 3 to the plots on the right side, it is obvious that the *RMSD* values are larger when the subgroup difference is large. For other conditions that are not provided in this figure, a similar pattern was observed.

In Figure 3, DTM is also plotted. Although specific DTM values are not provided in this paper, DTM is always larger than *REMSD* value regardless of the level of subgroup difference. However, when the *RMSD* values of the linear method are compared to DTM for the condition of large subgroup difference, two of the fourteen linking conditions (i.e., C3 and C4) result in larger *RMSD* values compared to DTM at some score points. C3 and C4 are linking conditions, in which the two tests differ in difficulty. As can be seen from Figure 3, when the linear method is used for Condition 3 when the subgroup difference is large, the *RMSD* values are larger than DTM at the lower and upper raw score ranges, which indicates that discrepancies in conversions resulting from the linear method for Condition 3 with larger subgroup difference may make some practical differences, that is, the difference in lower and upper raw score range is not ignorable difference.

To further confirm this, the difference in conversions between the total group and the two subgroups are plotted in Figures 4 and 5 for Conditions 3 and 5,

respectively, using the unsmoothed equipercentile and linear methods under the condition of large subgroup difference. The plots are based on the unrounded scores and it seems that the conversions for the total and each subgroup do not differ much when the unsmoothed equipercentile method is used. The difference in conversions between the total group and each subgroup is less than one raw-score point except when the linear method is used for Condition 3. As is expected from Figure 3, the discrepancies in conversions between the total and each subgroup are relatively large at the lower and upper raw score ranges when the linear method is used for Condition 3 when the subgroup difference is large.

4 Summary and Discussion

Concordance, which links tests that are built based on different content or statistical specifications, is likely to violate the group invariance property. However, performing concordance is sometimes required in practice since it can reduce testing cost, testing time, and also can make it possible to compare relatively performance across different tests. The main purpose of this paper was to examine the effects of various test characteristics on group invariance of four linking methods in a concordance context using a simulation study. A single group design was employed.

Multiple forms of a large-scale mathematics test were used to construct an item pool as a basis for creating various linking test forms. From the pool 14 pairs of concordance conditions which differ in test difficulty and test length were created. To examine the extent of group invariance, different concordance conditions were compared. Since two tests in each linking condition were different in terms of test difficulty or test length, the relationship between those two tests can be considered concordance rather than equating. The two tests in each concordance condition did not have any items in common and had relatively high correlations. Conversion tables were obtained for two subgroups as well as the total group using the linear, parallel linear, unsmoothed equipercentile, and postsmoothed equipercentile methods, and group invariance statistics were computed.

The results of this study are summarized as follows:

- A group with large difference in ability between subgroups shows more group dependency compared to other two groups where subgroups have no difference or small difference in ability, especially when two tests are different in terms of test length or test difficulty.

- When two tests are dissimilar in terms of test difficulty and test length, the assumption (i.e., a linear relationship between the two sets of scores) of the linear and parallel linear methods does not seem to be appropriate. A linear equating method assumes that the two forms to be linked have basically the same raw-score distribution (Buras, 1996). Yin et al. (2004) indicated that when two tests were different in content specifications and/or statistical properties, equipercentile methods were more

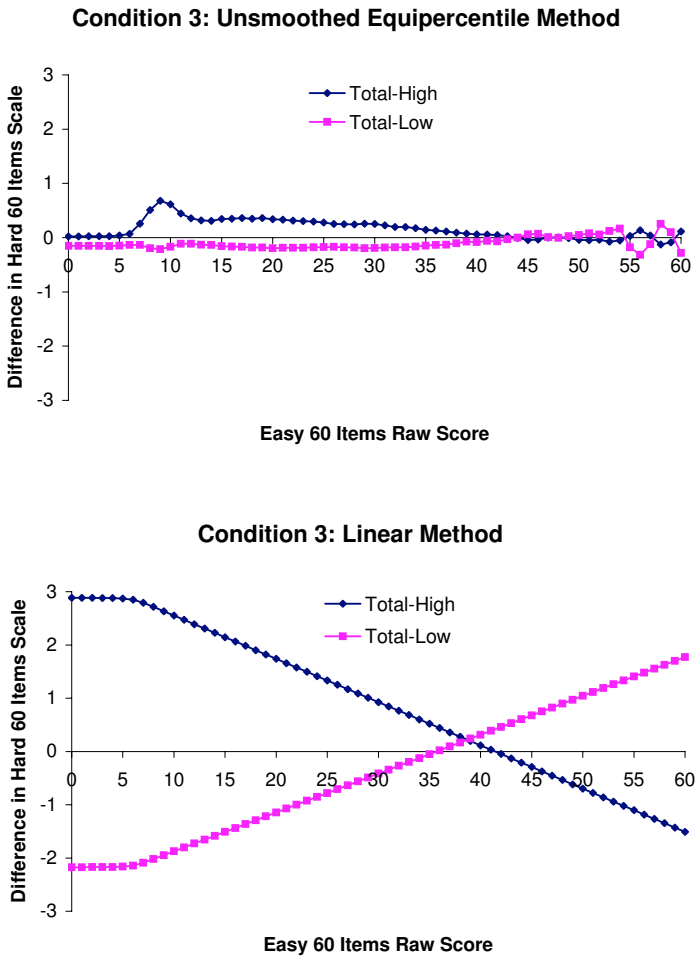

Figure 4: Difference in Conversions for Condition 3 between Linear and Unsmoothed Equipercentile Methods for Large Subgroup Difference

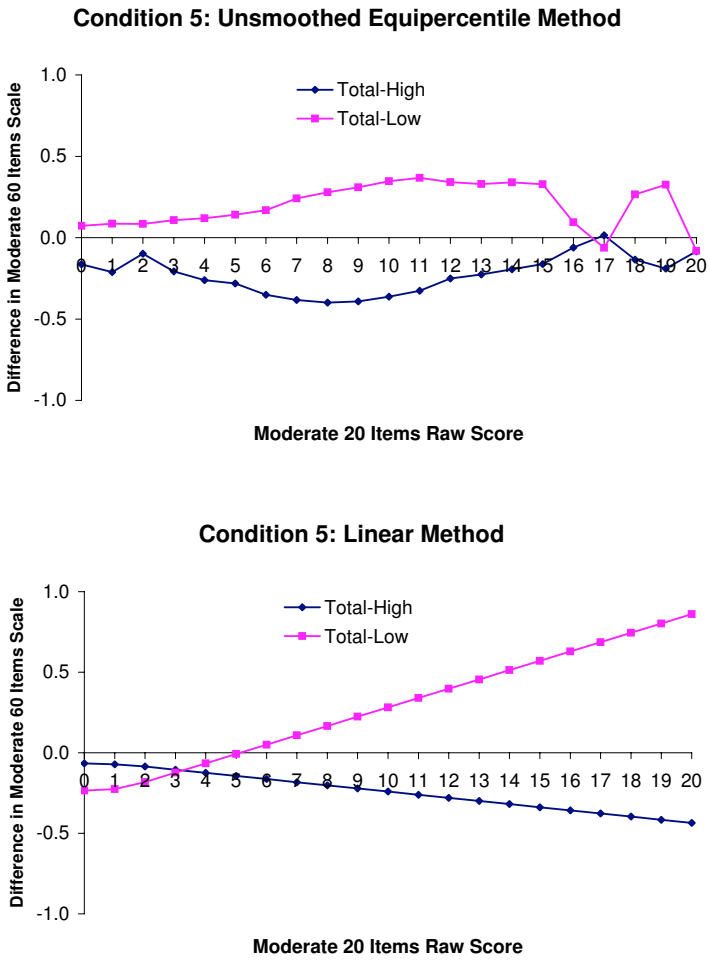

Figure 5: Difference in Conversions for Condition 5 between Linear and Unsmoothed Equipercentile Methods for Large Subgroup Difference

appropriate than linear methods. This study also shows that when two tests are different in terms of test difficulty and test length, the linear and parallel linear methods do not seem to be appropriate.

- The degree of group dependency for the linear method is relatively larger than that for the other three methods, especially when two tests to be linked are different in the difficulty level.
- The parallel linear method seems to be inappropriate for some linking conditions when scores on two tests are not linearly related. In addition, as Yin et al. (2004) stated if the group invariance property is not satisfied, the parallel linear method probably minimizes subgroup differences compared to the total group since the parallel linear method simplifies differences in conversions for subgroups into an intercept difference. Even though the parallel linear method might be somewhat inappropriate for some linking conditions, it shows relatively small degrees of group dependency across all conditions, and it works especially well when the subgroup difference was small.
- A linking pair consisting of tests with a large number of items (a long test) tends to be more group invariant than a linking pair consisting of tests with a small number of items (a short test).
- Linking scores on an easy test to those on a hard test is less group dependent than linking scores on a hard test to those on an easy test when equipercentile methods are used.
- Linking scores on a long test to those on a short test is less group dependent than linking scores on a short test to those on a long test.
- Sample size does not seem to influence group invariance as much as the number of items and test difficulty do.
- When test difficulty is dissimilar between tests to be linked, the resulting conversion is more group dependent than when test difficulty is similar.

Although not manipulated intentionally, the difference in reliability between two tests to be linked and the direction of linking seems to affect the degree of group invariance. Since the subgroups used in this study were defined by ability levels, a more reliable test would have separated two ability subgroups better than a less reliable test. That is, a more reliable test will be more subgroup sensitive, which implies that it will be more subgroup dependent. On the other hand, a relatively less reliable test will mask true differences that exist between different ability subgroups and the linking results will be less group dependent. When a relatively more reliable test is put on the scale of a relatively less reliable test, the linking results will be more group invariant (Brennan, personal communication, November 18, 2008). Even though the results of the present study were consistent with this speculation on the relationship between the

difference in reliability and the degree of group invariance, a more thorough examination seems necessary in future research.

Choice of a linking method seems to affect the differences in conversions obtained from different subgroups. Huh and Kolen's (2006) study showed that conversions obtained from different inking methods were different, which, in turn, resulted in different extents of group invariance. Yin et al. (2004) suggested that when content specifications and/or statistical properties of the two tests were significantly dissimilar, the transformation relationships using the linear method appeared to be group dependent. Hence, the choice of a linking method needs to be made with careful consideration of test and data characteristics and also the usage of the conversion results. The results of the present simulation study supported the conclusions drawn from previous studies, and also showed practical impact of performing concordance on group dependency using tests that differ in the difficulty level, length, and reliability.

In this study, two tests were built assuming IRT unidimensionality–that is, it was assumed that the two tests measured the same construct. In a practical concordance study, however, the construct measured by two tests will be similar but not strictly identical, and the correlation between the two ability vectors will differ from unity. Thus, a useful future study would involve a multivariate approach to constructing tests for concordance by incorporating different levels of correlation between two construct scores. This study could also be expanded to look at the content effect on group invariance by using a multidimensional approach or by re-sampling real data based on content. In addition, a subsequent simulation study could be conducted to examine the effect of a different definition of subgroups other than ability on group invariance.

5 References

Brennan, R. L. (2004). *Manual for LEGS (Version 2.0)* (CASMA Research Report No. 3). [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma/).

Buras, A. (1996, January). *Test equating procedures: A primer on the logic and applications of test equating.* Paper presented at the annual meeting of Southwest Educational Research Association, New Orleans, LA.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement, 28*, 227–246.

Dorans, N. J., & Feigenbaum, M. D. (1993). Equating issues engendered by changes to the new SAT I and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (Research memorandum 94–10). Princeton, NJ: Educational Testing Service.

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

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT Assessment and recentered SAT I sum scores. *College and University, 73*, 24–34.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of scoring across gender groups for three Advanced Placement Program examinations. In N. J. Dorans, (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS RR–03–27, pp. 79–118). Princeton, NJ: Educational Testing Service.

Ercikan, K. (1997). Linking statewide tests to the NAEP: Accuracy of combining test results across states. *Applied Measurement in Education, 10*, 145–159.

Hanson, B. A., Harris, D. J., Pommerich, M., Sconing, J. A., & Yi, Q. (2001, February). *Suggestions for the evaluation and use of concordance results* (ACT Research Report Series 2001–1).

Huh, N. R., & Kolen, M. J. (2006). *Group invariance in a concordance context.* Paper presented at the National Council on Measurement Education Annual Meeting, San Francisco.

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

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*, 129–140.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102.

Linn, R. L., & Kiplinger, V. L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education, 8*, 135–156.

Liu, J., Feigenbaum, M., & Dorans, N. (2005). *Invariance of Linkings of the Revised 2005 SAT Reasoning Test^{TM}to the SAT®I: Reasoning Test Across Gender Groups.* Princeton, NJ: Educational Testing Service.

McLaughlin, D. (1998). *Study of the linkages of 1996 NAEP and State Mathematics Assessments in four states.* Final Report. John C. Flanagan Research Center, Education Statistics Services Institute, American Institutes for Research. American Institutes for Research, Palo Alto, CA.

Mislevy, R. J. (1992). *Linking educational assessments: Concept, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement, 28*, 247–273.

Segall, D. O. (1997). Chapter 19. Equating the CAT-ASVAB. In W. A. Sands, B. K. Walters, & J. R. McBrade (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181–198). Washington, DC: American Psychological Association.

Tateneni, K., & Dorans, N. J. (2003). Invariance of linkages for the free-response, multiple-choice and composite scores on Advanced Placement Program examinations. In N. J. Dorans, (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS RR–03–27, pp. 37–56). Princeton, NJ: Educational Testing Service.

Williams, V., Billeaud, K., Davis, L., Thissen, D., & Sanford, E. (1995). *Projecting to the NAEP scale: Results from the North Carolina End-of-Grade Testing Program* (Technical report #34). Chapel Hill, NC: National Institute of Statistical Sciences, University of North Carolina, Chapel Hill.

Yi, Q., Harris, D. J., Gao, X. (2007). *Invariance of equating functions across different subgroups of examinees taking a science achievement test*. Manuscript submitted for publication.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement, 28*, 274–289.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items* [Computer program]. Chicago: Scientific Software International.