Listening. Learning. Leading.®

Research Report
ETS RR–14-41

The Invariance of Latent and Observed Linking Functions in the Presence of Multiple Latent Test-Taker Dimensions

Neil J. Dorans

Peng Lin

Wei Wang

Lili Yao

December 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Invariance of Latent and Observed Linking Functions in the Presence of Multiple Latent Test-Taker Dimensions

Neil J. Dorans, Peng Lin, Wei Wang, & Lili Yao

Educational Testing Service, Princeton, NJ

This study examines linking relationships among latent test scores and how these latent linking relationships relate to observed-score linkings. Equations are used to describe the effects of correlation between underlying latent dimensions and the similarity or dissimilarity of test composition on linking functions among latent test scores. These equations describing relationships among latent test scores are used to model the results obtained from a previous simulation study, which illustrated that if the two tests have parallel structure then the linking relationship between their observed scores is subpopulation invariant regardless of the correlations between the underlying latent dimensions. The equations also model the effect that the degree of correlation between the latent dimensions has on equatability as the structure departs from parallelism.

Keywords Multidimensionality; simple structure; linking; invariance latent variable; observed score

doi:10.1002/ets2.12041

Dorans and Lawrence (1999) maintained that the dimensionality detected in relationships among item scores is not necessarily the same as the dimensionality observed among test scores. They used data from the Dorans and Lawrence (1987) investigation of the factors in the *SAT*® data to illustrate this point. They advocated that the choice of dimensionality technique should be based on the purpose of the dimensionality analysis.

The psychometric model employed in the Dorans and Lawrence (1987, 1999) studies was the common factor model (Mislevy, 1986; Mulaik, 1972; Thurstone, 1947). According to that latent variable model, a *common factor* is a hypothetical variable that contributes to the variance of two or more observed variables. In addition to common factors, each observed variable has one *unique factor*. Each hypothetical unique factor contributes to the variance of only one of the observed variables.

When item score data are considered, there are as many unique factors as there are items. In addition, there are the common factors. When we collapse the item data into a single composite score, we lose access to item-level dimensionality and are left with a single test score dimension. Only one observable is present. When there is only one observable, there is only one ordering of test takers, one factor; common factors need multiple variables to emerge.

We use dimension instead of factor because it has less surplus meaning than factor has. Sometimes the word *factor* is presumed to be an attribute of test takers that exists independently of the data. Dimension is less laden with that meaning. To better understand the implications of this dimensionality reduction for testing contexts, consider that the common factor model posits that the reliable variability of an item score on a test is influenced by systematic sources shared with other items in the test as well as by a reliable source of variation that is unique to that item and independent of other items in the test. These specific components, which differ from measurement error in that they are systematic influences on test-taker performance, may be related to item content, item location, or other conditions of measurement. From this perspective, at least NI (number of items) dimensions are at play in item score data: one unique dimension for each item plus the number of shared dimensions. When item scores are reduced to a single composite score, there is only one score for each test taker. Hence only one dimension exists, albeit a potentially complex one.

Dorans and Lawrence (1999) made a distinction between item-level dimensionality and test-level dimensionality (and subtest dimensionality). Test score equating does not require that all items measure the same single dimension. It simply requires that the scores to be equated measure the same dimension, even if it is a complex one. The restrictive assumption

Corresponding author: N. J. Dorans, E-mail: ndorans@ets.org

of item unidimensionality, that all items measure the same dimension, is not required to equate scores at the level of a total test.

Lin and Dorans (2011) investigated subpopulation invariance in a simulated scenario related to vertical linking. Their hypothetical subpopulations could be thought of as defined by the grade of examinees, where the ability distribution of subpopulations varies across grades. That study attempted to simulate what could happen in a vertical scaling with a shift in content structure, a change in test difficulty, and differential shifts in students' ability proficiency across the dimensions underlying performance on the content domains. Specifically, Lin and Dorans assumed that two distinct content domains were taught and tested across three grade levels. While performance in each content domain was presumed to be a function of a single psychometric construct dimension, the test measured two dimensions. At each grade level, proficiencies on the two dimensions were simulated to be highly related and less related.

Parallel structure exists when the proportions of subsets of items that measure different dimensions are the same across the tests to be linked and the relationships of the items to the underlying dimensions are also the same. In essence, this is a simple structure (Thurstone, 1947) in which each subset measures only one dimension, but different subsets measure different dimensions. A brief description of the Lin and Dorans (2011) simulation design study is provided in the section of this article titled "Illustration with Data Generated by Lin and Dorans (2011)."

Lin and Dorans (2011) demonstrated that when content structure is not parallel, that is, the proportion of items measuring each of the two dimensions differed across the tests to be linked, subpopulation invariance of equating functions, one of the requirements of equating (Holland & Dorans, 2006; Lord, 1980), is not achieved. In addition, they found that the degree to which it can be achieved depends on the correlation between the dimensions underlying performance on the content domains. The results from the study suggested that when there is a construct shift across tests to be linked, subpopulation invariance should not be assumed without further investigation about the characteristics of the tests and the subpopulations to which the linking functions are applied.

Lin and Dorans (2011) also found that subpopulation invariance of observed-score equating can be achieved with tests that are composed of subsets of items that measure different dimensions, provided that the tests are parallel in content structure. They also found that this invariance holds whether the correlation between the dimensions underlying the test performance is weak or strong. This finding confirmed that violations of unidimensionality at the level of items scores need not present problems for observed-score equating, provided that the simple structures of each test are equivalent. Another way of saying this is that test score equating may be robust to violations of unidimensionality at the level of item scores, provided that the assumption of unidimensionality is met on the test score level through careful content balancing that produces parallel test forms.

The Lin and Dorans (2011) study used a multidimensional item response theory (MIRT) model (Reckase, 1985, 2009) to generate simulated data and focused on linking functions that would be used with observed test scores. They examined observed-score equating methods. In this article, we examine linking relationships among latent test scores. In particular, this study uses analytic relationships among latent variables to better understand the Lin and Dorans results.

In the section titled "Modeling the Latent Space Presumed to Underlie Observed Performance," we describe a latent variable model that is presumed to underlie observed test performance. It shows how performance on latent dimensions that are item-free can be translated to performance on latent variables associated with two tests that are composed of items that measure one of these latent dimensions. The section titled "Latent Linking in a Single Population" contains the mathematics for the linear linking of latent variables underlying test performance. In the section titled "Illustration with Data Generated by Lin and Dorans (2011)," we describe the Lin and Dorans simulation study design. The final two sections present the linear latent linkings alongside linkings taken from Lin and Dorans, explain the findings, and look toward future research.

Modeling the Latent Space Presumed to Underlie Observed Performance

In this section, we introduce the notation and terminology used in this study. Following that, we derive a series of mathematical expressions using this notation. It is important to note that for the most part the mathematical expressions are about entities in the latent space instead of in the domain of observables.

Table 1 Summary of Notation (General)

Symbol	Description
Q	Population of test takers
NLD	Number of latent dimensions
NI	Number of items
θ	NLD-by-1 vector of examinee's ability
i	An index for an item in Form X or Y
$\mathbf{a}_i$	NLD-by-1 vector of dimension weights on item i
$\mathbf{A}_x$	NLD-by-NI matrix of dimension weights by item for Test Form X
$\mathbf{A}_y$	NLD-by-NI matrix of dimension weights by item for Test Form Y
BIS_i	Binary item score for item i
CLIS_i	Continuous latent item score for item i
CLIS_{xj}	Continuous latent item score for item j on Form X
CLIS_{yk}	Continuous latent item score for item k on Form Y
$\mathbf{d}_i$	Difficulty of item i
$\mathbf{d}_x$	NI-by-1 vector of item difficulties for Form X
$\mathbf{d}_y$	NI-by-1 vector of item difficulties for Form Y
$\text{ntcpt}(x \rightarrow y)$	Intercept term in the linear linking function when scores on Form X are linked back to Form Y scale
$\text{slp}(x \rightarrow y)$	Slope term in the linear linking function when scores on Form X are linked/equated back to Form Y scale
LT_x	Latent test score on Form X
LT_y	Latent test score on Form Y
$\mathbf{Ave}_q$	Mean of variables (e.g., item score, test score, ability) of Population Q
Var_q	Variance of variables (e.g., item score, test score, ability) of Population Q
$\mathbf{Cov}_q(\boldsymbol{\theta})$	Covariance matrix for $\boldsymbol{\theta}$ of Population Q
Corr_q	Correlation between latent test scores in Population Q

Notation and Latent Entities

Let's presume that for each test taker, underlying his or her observed performance on a test, O, there is a bounded[1] version of a continuous latent variable, T, that is the expected value of performance on that test for individuals just like that test taker. According to the classical test theory, observed score (O) can be decomposed to the bounded true score (T) and error term (e), which can be expressed as $O = T + e$. An unbounded version of true score is introduced in the next section. This unbounded true score (LT) is presumed to be a one-dimensional linear function of test-taker ability in the latent space, which can be multidimensional. The unbounded true score is referred to as the latent test score in the latent space, whereas the bounded true score is referred to as the classical test theory definition of a true score. In the rest of this article, we focus on the unbounded latent test score, LT.

Let X and Y denote two forms of the same or different multiple-choice only test(s). In each test form, it is assumed that some number of constructs or latent dimensions (NLD) C1, C2, . . . , and $\text{C}_{\text{NLD,}}$ are measured. Each construct is related to proficiency in that content domain. In addition, there are NI items in each test form.

The linear linkage between latent test scores for Forms X and Y is studied in the latent space among different subpopulations. Latent test scores on Form X are linked back to the latent test score scale on Form Y. Table 1 summarizes the various symbols that are introduced and used in this section and the next section.

Latent Test Scores

This subsection illustrates how latent test scores for both Forms X and Y are a function of the underlying construct dimensions. Assume that members of a single population, Q, take both Form X and Form Y. For simplicity, assume that each item in the two forms measures only one dimension. Both forms, however, contain items that measure different dimensions. In addition, there is a multidimensional space of NLD latent construct values that accounts for a portion of predictable item and test performance. For example, NLD is 2 when the test form is composed of items that measure either math or reading. Furthermore, in each form, there are NI latent item scores, one for each item. These latent item scores can be expressed as linear combinations of the NLD latent dimensions. The latent test score for either test form is simply the sum of the latent item scores for items on that form. Hence, the latent test scores are also linear combinations of the latent construct dimensions.

Assume that a commonly used MIRT model holds. Here the predictable part of an item score can be expressed as a function of the NLD latent dimensions:

$$P_q\left(\mathrm{BIS}_i = 1\,|\boldsymbol{\theta}\right) = \frac{e^{\left(\mathbf{a}_i'\boldsymbol{\theta}-\mathrm{d}_i\right)}}{1+e^{\left(\mathbf{a}_i'\boldsymbol{\theta}-\mathrm{d}_i\right)}}. \tag{1}$$

Here $P_q(\mathrm{BIS}_i = 1|\boldsymbol{\theta})$ represents the probability of obtaining a binary item score on item i of 1 as a function of NLD-dimensional $\boldsymbol{\theta}$, expressed as NLD-by-1 vector, where the subscript q stands for a test taker from subpopulation Q. Hence, $P_q(\mathrm{BIS}_i = 1|\boldsymbol{\theta})$ is just the item true score on item i for test-taker q.

In addition, for test-taker q, there is still a latent score for each item, which can be expressed as a linear function of the abilities on each dimension. Note that every latent item score is continuous, instead of being binary, in the latent space. Taking the log odds of Equation 1 yields,

$$\mathrm{CLIS}_i = \ln\left(P_q\left(\mathrm{BIS}_i = 1\,|\boldsymbol{\theta}\right)/P_q\left(\mathrm{BIS}_i = 0\,|\boldsymbol{\theta}\right)\right)_i = \mathbf{a}_i'\boldsymbol{\theta} - \mathrm{d}_i, \tag{2}$$

where CLIS_i is defined as the continuous latent item score on item i (see Table 1). In the equations above, $\mathbf{a}_i'$ is a 1-by-NLD vector, specifying the weights of the NLD dimensions on item i, and d_i is a parameter that is related to item difficulty.

In addition to the test true score (referred to as bounded true score in the previous subsection), obtained by summing the NI $P_q(\mathrm{BIS}_i = 1|\boldsymbol{\theta})$, there is also a continuous latent test score (referred to as unbounded true score in the previous subsection) obtained by summing the continuous latent item scores (CLIS_i).

For Form X with item $j = 1, 2, \ldots,$ NI, the latent test score is:

$$\mathrm{LT}_x = \sum_j^{\mathrm{NI}} \mathrm{CLIS}_{xj} = \sum_j^{\mathrm{NI}} (\mathbf{a}_{xj}'\boldsymbol{\theta} - \mathrm{d}_{xj}) = \mathbf{1}'\mathbf{A}_x'\boldsymbol{\theta} - \mathbf{1}'\mathbf{d}_x, \tag{3}$$

where $\mathbf{A}_x{}'$ is a NI-by-NLD matrix of item by dimension weights (related to discrimination power in the IRT models), $\mathbf{1}'$ is a 1-by-NI vector of ones, and $\mathbf{d}_x$ is a NI-by-1 vector of item difficulties for Form X.

The average score of LT_x, $\mathrm{Ave}_q(\mathrm{LT}_x)$, and the variance of LT_x, $\mathrm{Var}_q(\mathrm{LT}_x)$, in subpopulation Q can be expressed as:

$$\mathrm{Ave}_q\left(\mathrm{LT}_x\right) = \sum_j^{\mathrm{NI}} \mathrm{Ave}_q(\mathrm{CLIS}_{xj}) = \mathrm{Ave}_q\left(\mathbf{1}'\mathbf{A}_x'\boldsymbol{\theta}\right) - \mathrm{Ave}_q\left(\mathbf{1}'\mathbf{d}_x\right) = \mathbf{1}'\mathbf{A}_x'\mathbf{Ave}_q\left(\boldsymbol{\theta}\right) - \mathbf{1}'\mathbf{d}_x, \tag{4}$$

and

$$\mathrm{Var}_q\left(\mathrm{LT}_x\right) = \mathbf{1}'\mathbf{A}_x'\mathbf{Cov}_q\left(\boldsymbol{\theta}\right)\mathbf{A}_x\mathbf{1}. \tag{5}$$

Similar to Form X, the continuous latent test score LT_y for Form Y based on the sum of items $k = 1, 2, \ldots,$ NI can be expressed as:

$$\mathrm{LT}_y = \sum_k^{\mathrm{NI}} \mathrm{CLIS}_{yk} = \sum_k^{\mathrm{NI}} \left(\mathbf{a}_{yk}'\boldsymbol{\theta} - \mathrm{d}_{yk}\right) = \mathbf{1}'\mathbf{A}_y'\boldsymbol{\theta} - \mathbf{1}'\mathbf{d}_y, \tag{6}$$

with mean and variance,

$$\mathrm{Ave}_q\left(\mathrm{LT}_y\right) = \sum_k^{\mathrm{NI}} \mathrm{Ave}_q(\mathrm{CLIS}_{yk}) = \mathrm{Ave}_q\left(\mathbf{1}'\mathbf{A}_y'\boldsymbol{\theta}\right) - \mathrm{Ave}_q\left(\mathbf{1}'\mathbf{d}_y\right) = \mathbf{1}'\mathbf{A}_y'\mathbf{Ave}_q\left(\boldsymbol{\theta}\right) - \mathbf{1}'\mathbf{d}_y, \tag{7}$$

and

$$\mathrm{Var}_q\left(\mathrm{LT}_y\right) = \mathbf{1}'\mathbf{A}_y'\mathbf{Cov}_q\left(\boldsymbol{\theta}\right)\mathbf{A}_y\mathbf{1}. \tag{8}$$

Latent Linking in a Single Population

Again, in the previous section, the latent test scores LT_x and LT_y are obtained for Forms X and Y, respectively. In this section, the linking function is derived to link the latent test scores on Form X back to the latent test score scale on Form Y.

Only the linear linking method is considered in this study. The slope in the linear linking function is $\sqrt{\text{Var}_q\left(\text{LT}_y\right)}/\text{Var}_q\left(\text{LT}_x\right)$, which can be expressed as:

$$\text{slp}\left(x \rightarrow y\right) = \sqrt{\mathbf{1'A}_y'\mathbf{Cov}_q\left(\boldsymbol{\theta}\right)\mathbf{A}_y\mathbf{1}/\mathbf{1'A}_x'\mathbf{Cov}_q\left(\boldsymbol{\theta}\right)\mathbf{A}_x\mathbf{1}}. \tag{9}$$

The intercept can be computed the following formula:

$$\text{ntcpt}\left(x \rightarrow y\right) = \left(\mathbf{1'A}_y'\mathbf{Ave}_q\left(\boldsymbol{\theta}\right) - \mathbf{1'd}_y\right) - \text{slp}\left(x \rightarrow y\right)\left(\mathbf{1'A}_x'\mathbf{Ave}_q\left(\boldsymbol{\theta}\right) - \mathbf{1'd}_x\right). \tag{10}$$

Note that when $\mathbf{A}_x$ and $\mathbf{A}_y$ are identical, which means parallel structures for Form X and Form Y, the slope is 1, regardless of the covariance among the fundamental latent dimensions, the number of dimensions, or the difficulties of the items comprising the tests. If the slope is 1, the intercept is simply the difference between the difficulties of these two forms.

In addition, the correlation between the latent test scores LT_x and LT_y in Q can be computed using the equation as below:

$$\text{Corr}_q\left(\text{LT}_x, \text{LT}_y\right) = \left(\mathbf{1'A}_x'\mathbf{Cov}_q\left(\boldsymbol{\theta}\right)\mathbf{A}_y\mathbf{1}\right)/\sqrt{\left(\mathbf{1'A}_x'\mathbf{Cov}_q\left(\boldsymbol{\theta}\right)\mathbf{A}_x\mathbf{1}\right)\left(\mathbf{1'A}_y'\mathbf{Cov}_q\left(\boldsymbol{\theta}\right)\mathbf{A}_y\mathbf{1}\right)}. \tag{11}$$

Note that when $\mathbf{A}_x$ and $\mathbf{A}_y$ are identical, the correlation between LT_x and LT_y in Q is 1, regardless of the number of the fundamental dimensions that underlie the latent space, or the magnitude of the correlations among the dimensions. This is a consequence of parallel structure. Dorans and Lawrence (1999) made this point while noting distinctions among item-level dimensionality, testlet-level dimensionality, and total-test dimensionality. Test score equating does not require that items measure a single dimension. It simply requires that test scores measure the same dimension in the same way even if the dimension is a complex one. The very restrictive assumption of item unidimensionality is not required to equate scores at the level of a total test. This point will become apparent as we use the mathematics above to "explain" the results obtained by Lin and Dorans (2011).

Illustration with Data Generated by Lin and Dorans (2011)

In Lin and Dorans' study (2011), it was assumed that two distinct content domains were taught and tested across three grade levels: Grade L, Grade M, and Grade H. At each grade level, proficiencies on the two dimensions might be highly related, such as algebra and geometry, or less related, such as math and reading. Each item in the tests measures only one content domain: either C1 or C2. Q_{L}, Q_{M}, and Q_{H} are the test taking populations in Grades L, M, and H, respectively. Table 2 summarizes the specific notation for the simulation study; detailed explanations are provided later.

Simulated Tests

In Lin and Dorans (2011), nine simulated tests were developed by crossing three levels of difficulty with three levels of content specifications. The three difficulty levels were easier (e), moderate (m), and harder (h). When the test form was an easier one, the mean of difficulty parameter is 0 for items in each of the two dimensions. For a form with moderate difficulty, the means of difficulty parameter for the items in C1 and C2 were 0.15 and 0.25, respectively. For the harder form, the means of difficulty parameter for the C1 items and the C2 items were both 0.30. The content specifications differed with respect to the number of items measuring C1 and C2, respectively, in a test. For the three levels of content specification, the ratio of items measuring C1 to those measuring C2 were 4:1, 1:4, and 3:2.

Crossing the three difficulty levels with the three content specifications yields nine tests, as indicated in Table 3. Each of these simulated nine tests was administered along with a Form X_{e}(4:1), which was parallel to Y_{e}(4:1), in each of three subpopulations: Q_{L}, Q_{M}, and Q_{H}. The tests contained 80 items.

Subpopulation Characteristics

Table 4 contains the means and standard deviations of the abilities underlying performance on C1 and C2 for subpopulations Q_{L}, Q_{M}, and Q_{H}. Note that differences between Q_{L} to Q_{H} are comparable on the dimensions underlying performance

Table 2 Summary of Notation (Specific)

Symbol	Meaning
Q_L	Subpopulation L
Q_M	Subpopulation M
Q_H	Subpopulation H
C1	1st content domain
C2	2nd content domain
$\boldsymbol{\theta}_1$	Examinee's ability on C1
$\boldsymbol{\theta}_2$	Examinee's ability on C2
$\mathbf{a}_i$	2-by-1 vector of dimension weights for item i
$\mathbf{a}_{i1}$	Dimension weight on C1 for item i
$\mathbf{a}_{i2}$	Dimension weight on C2 for item i
$\mathbf{d}_i$	Item difficulty for item i
M1: (C1:C2 = 4:1)	1st type of content specification
M2: (C1:C2 = 1:4)	2nd type of content specification
M3: (C1:C2 = 3:2)	3rd type of content specification
X (C1:C2)	Test Form X with content mix of (C1:C2)
Y (C1:C2)	Test Form Y with content mix of (C1:C2)
Y_e	Easier difficulty version of Test Form Y
Y_m	Moderate difficulty version of Test Form Y
Y_h	Harder difficulty version of Test Form Y
$\boldsymbol{\mu}$	2-by-1 vector of mean of examinee ability $\boldsymbol{\theta}$
$\boldsymbol{\psi}$	2-by-2 variance-covariance matrix of examinee ability $\boldsymbol{\theta}$

Table 3 Combinations of Content Specification and Difficulty for the Nine Tests Linked to Form X_e(4:1)

Content specification	Difficulty levels of Y Easier (e)	Moderate (m)	Harder (h)
M1: (4:1)	Y_e(4:1)	Y_m(4:1)	Y_h(4:1)
M2: (1:4)	Y_e(1:4)	Y_m(1:4)	Y_h(1:4)
M3: (3:2)	Y_e(3:2)	Y_m(3:2)	Y_h(3:2)

Table 4 The Ability Distribution of the Three Subpopulations

Subpopulation	Mean(C1, C2)	Standard deviation(C1, C2)
Subpopulation Q_L	(0, 0)	(1, 1)
Subpopulation Q_M	(0.15, 0.25)	(1, 1)
Subpopulation Q_H	(0.30, 0.30)	(1, 1)

on C1 and C2, with an increase of 0.30 standard deviation units on each. In contrast, Q_M is half way between Q_L and Q_H on C1, but 0.25 above Q_L and 0.05 below Q_H on C2. The standard deviation of the abilities underlying performance on C1 and C2 is 1 for all three subpopulations.

In addition, Lin and Dorans (2011) also varied the correlation between the abilities underlying performance on C1 and C2 within each subpopulation to better examine the effects of multidimensionality. The four levels of correlations between abilities underlying performance on C1 and C2 (not observed scores) were 0.30, 0.50, 0.70, and 0.95. In this study, we only report results for correlation levels of 0.50 and 0.95. Table 5 summarizes the factors and their levels considered in this study. Because the factors are crossed, there are, in total, 54 conditions or 54 linking functions.

Data Generation

In the Lin and Dorans (2011) study, the probability of a simulee correctly answering item i was computed based on the 1PL-MIRT model,

$$P_s\left(\mathrm{BIS}_i = 1|\theta_s\right) = \frac{e^{\left(\mathrm{a}_i'\theta_s - d_i\right)}}{1 + e^{\left(\mathrm{a}_i'\theta_s - d_i\right)}},$$

Table 5 Factors of Investigation

Factor	Level
Test difficulty (only for Form Y)	3 (easier, moderate, harder)
Test content specification (only for Form Y)	3 (C1:C2 = 4:1, 1:4, and 3:2)
Correlation between abilities underlying performance on C1 and C2	2 (0.50 and 0.95)
Subpopulation ability	3 (mean: [0, 0], [0.15, 0.25], and [0.30, 0.30]; standard deviation: [1, 1])

where s denotes a simulee and d_i is a scalar denoting the difficulty of item i. θ_s, a 2-by-1 vector, denotes the simulee's ability. a_i is a 2-by-1 vector and denotes dimension weights for item i. As described earlier, each item only measured one dimension (C1 or C2); therefore, a_i is either (1, 0) when the items measured the dimension defined by C1 only or (0,1) when the items measured solely C2 dimension. The parameter values of a and d for the items in the easy tests with content mixes C1:C2 of 4:1, 1:4, and 3:2 are provided in Tables A1, A2, and A3, respectively. For test forms with moderate difficulty, 0.15 and 0.25 were added to d values of C1 items and C2 items on the easy form with comparable structure, respectively. For the harder forms, 0.30 was added to d values of all the items on the easy form with comparable structure.

Under each combination of conditions, for each of the three subpopulations, the item responses for 100,000 simulees were generated. To link scores on Form X to score scale of Form Y, single group linear and equipercentile linking were carried out.

The Analytic Predictions

In addition to the simulated results from the Lin and Dorans (2011) study just described, we produced analytical predictions based on the model described in the two previous sections. Specifically, we converted the items level parameters for slope and intercept, described in the Appendix and the text associated with Tables 3 and 5, and the population parameters related to means and covariance among the NLD underlying dimensions, depicted in Tables 4 and 5, to estimates of performance on the unbounded latent tests scores, LT_x and LT_y.

We then used the linear conversion defined by Equations 9 and 10 for linking latent test scores, LT_x and LT_y. It is important to note that a linear scale transformation was conducted to put the unbounded LT_x and LT_y scales on the bounded observed-score scales for the X_e test. For example, LT_x was the result of a linear transformation from the underlying latent dimensions. Then these scores, in effect, were linearly transformed to the scale of the simulated observed scores on Form X_e in each subpopulation (Q_L, Q_M, Q_H) by giving them the same mean and standard deviation as scores on X_e in each subpopulation. The same occurred for LT_y except by LT_y giving the same mean and standard deviation as observed scores on each Y form in each subpopulation. These scalings were necessary to permit direct comparisons between the analytical results and the simulated results, which were in the metric of the observed scores on X_e.

Results

Figure 1, and each subsequent figure, contains four panels. Each curve in each panel presents a difference in linking functions between what was obtained by a particular method and what is expected when equating two strictly parallel forms (i.e., parallel in construct measured and difficulty), namely the identity function in each of three subpopulations. In the plots, the horizontal axis is the total score on Form X and the vertical axis is the difference between the linking functions and the identity line.

Each of the six figures is a result of crossing three levels of content structure or relative weight given to C1 versus C2 and two levels of correlation between the latent variables underling performance on C1 and C2. Figures 1, 3, and 5 contain the results for the nearly unidimensional case where the fundamental latent variables correlate 0.95. Figures 2, 4, and 6 contain the results for the clearly two-dimensional case where the fundamental latent variables correlate 0.50.

Figures 1 – 6 contain three sets of difference curves: one for equating X_e to Y_e, one for equating X_e to Y_m, one for equating X_e to Y_h. The abscissa is the raw score on X. The ordinate is the difference between the number of raw score points needed to make a score on X_e equivalent to a score on each Y and what we would expect if Form X_e were strictly parallel in difficulty to each version of Y, which would be no adjustment at all.

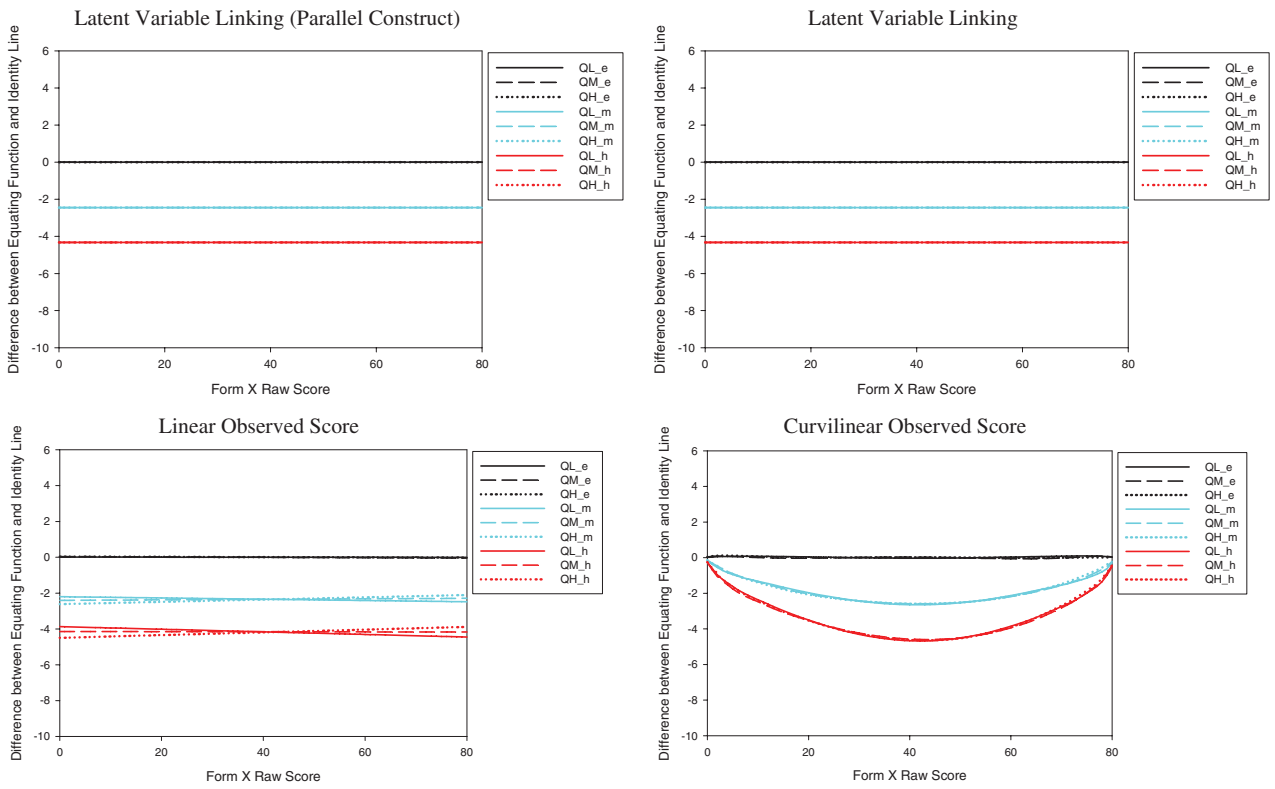

Figure 1 Linking $X_e(4{:}1)$ to $Y(4{:}1)$ when $r = 0.95$.

The upper left panel in each figure contains differences between the linear conversion defined by Equations 9 and 10 for linking latent test scores, LT_x and LT_y. As noted above, a linear scale transformation was conducted to put the unbounded LT_x and LT_y scores on the bounded observed-score scales in each subpopulation (Q_L, Q_M, and Q_H). Note there are three sets of horizontal lines. Each line represents a difference between a linking of latent test scores on tests that are parallel in content structure but which may differ in difficulty and the identify function.

Each of the three horizontal lines in the upper left panel represents three different lines, one for each subpopulation, Q_L, Q_M, Q_H. In other words, lines QL_e, QM_e, and QH_e are all coincident because the equating of $X_e(4{:}1)$ to $Y_e(4{:}1)$ is invariant across populations. The same holds for $X_e(4{:}1)$ to $Y_m(4{:}1)$ and $X_e(4{:}1)$ to $Y_h(4{:}1)$, which are represented by the lines QL_m, QM_m, and QH_m, and QL_h, QM_h, and QH_h, respectively.

The line of zero difference occurs when linking test forms of parallel content structure and equal difficulty, for example, when linking $X_e(4{:}1)$ to $Y_e(4{:}1)$. This zero difference line indicates that there is no need to adjust scores on $X_e(4{:}1)$ to make them equivalent to $Y_e(4{:}1)$. The line with a difference slightly below an ordinate value of −2 occurs when equating tests of parallel content but different difficulty, for example, when linking $X_e(4{:}1)$ to $Y_m(4{:}1)$. This difference line indicates that the there is a need to adjust scores on $X_e(4{:}1)$ by over two points to make them equivalent to $Y_m(4{:}1)$. The line with a difference below an ordinate value of −4 occurs when linking test forms of parallel content but even greater differences in difficulty, for example, when equating $X_e(4{:}1)$ to $Y_h(4{:}1)$. This difference line indicates that the there is a need to adjust scores on $X_e(4{:}1)$ by over four points to make them equivalent to $Y_h(4{:}1)$.

The line slightly below an ordinate value of −2 in Figures 1, 3, and 5 is observed when linking $X_e(4{:}1)$ to $Y_m(4{:}1)$ in a subpopulation where the latent variables underlying C1 and C2 correlate 0.95. When the latent variables underlying C1 and C2 correlate 0.50, this difference remains larger than −2 but by a slightly smaller amount, as seen in Figures 2, 4, and 6. When the correlation is 0.95, the horizontal line is slightly below −4, as seen in Figures 1, 3, and 5. The line slightly above an ordinate of −4 in the upper left panel of Figures 2, 4, and 6 represents linking $X_e(4{:}1)$ to $Y_h(4{:}1)$ when the correlation is 0.50. As the difference in test difficulty increases, larger amounts of scores units are needed to adjust raw scores on X to make them equivalent to raw scores on the Y. Lower correlations between the two fundamental latent variables attenuate the effect of the difficulty difference.

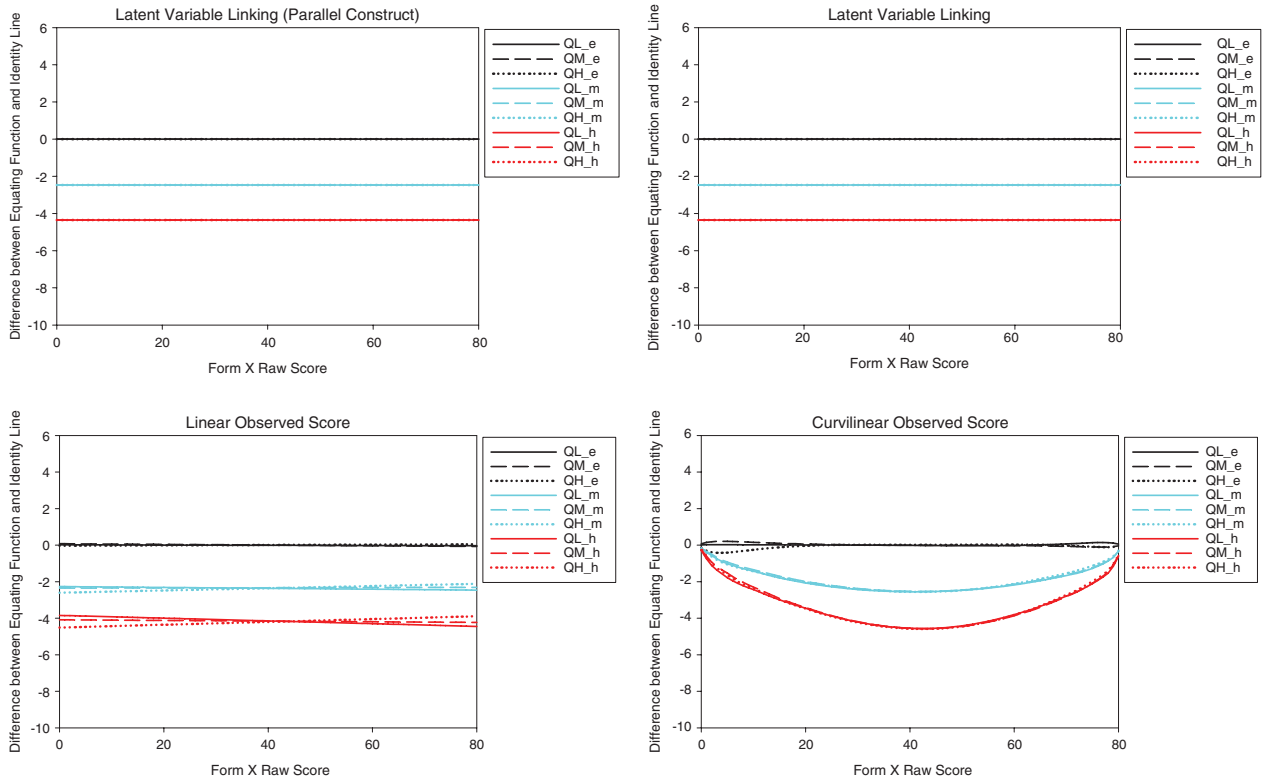

Figure 2 Linking $X_e(4{:}1)$ to $Y(4{:}1)$ when $r = 0.50$.

The rest of each figure contains for the particular combination of latent variable correlation (0.95 or 0.50) and content structure (see Table 5) the latent total score linking (upper right), the linear observed-score linking (lower left), and the equipercentile linking (lower right) for each of the three subpopulations (described in Table 4).

Subpopulation Invariance When Linking Tests Are Composed of Two Dimensions

The linking functions for $X_e(4{:}1)$ to $Y_e(4{:}1)$ were invariant across all three subpopulations described in Table 4. The difference curves for the latent variables linkings in these subpopulations in Figure 1 are coincident. This was also true for the linkings of $X_e(4{:}1)$ to $Y_m(4{:}1)$ and $X_e(4{:}1)$ to $Y_h(4{:}1)$. This invariance across subpopulations was observed for both correlations of 0.50 and 0.95, as can be seen in the upper left panel of all six figures.

In Figure 1, the correlation between fundamental latent variable is 0.95, and the content structures of $Y(4,1)$ and $X_e(4,1)$ are the same, but there are three levels of difficulty for $Y(4,1)$, represented by the subscripts, e, m, and h. Consequently, the upper right panel is identical to the upper left panel, where three parallel lines, one for each level of difficulty for Y, each represent three subpopulation invariant linking. The lower left panel contains the linear observed-score linking, and it is identical in shape for the linking of $X_e(4,1)$ to $Y_e(4,1)$ but deviates from a horizontal line when Y differs in difficulty from X, with the direction of deviation depending on the subpopulation.

The equipercentile linking functions of observed score for $X_e(4{:}1)$ to the three versions of $Y(4{:}1)$ are depicted in the lower right panel of Figure 1, where each curve plotted represents the difference between the linking functions and the identity line (expected with parallel forms) for the three subpopulations (Q_L, Q_M, and Q_H) under the three different difficulty levels (e, m, h). In Figure 1, the linking of $X_e(4{:}1)$ to $Y(4{:}1)$ looks the same across all three subpopulations for all three difficulty levels of Y. This indicates that subpopulation invariance holds for the equipercentile linking of $X_e(4{:}1)$ to each $Y(4{:}1)$, which are equatings, under all conditions.

Another finding in Figure 1 is that the equipercentile linking function of $X_e(4{:}1)$ to $Y_e(4{:}1)$ is equivalent to the identity line. But as Form $Y(4{:}1)$ gets harder, the discrepancy between the linking function and the identity line becomes larger, as expected. This confirms that equating results in an adjustment for the difference in difficulty between tests. However, the

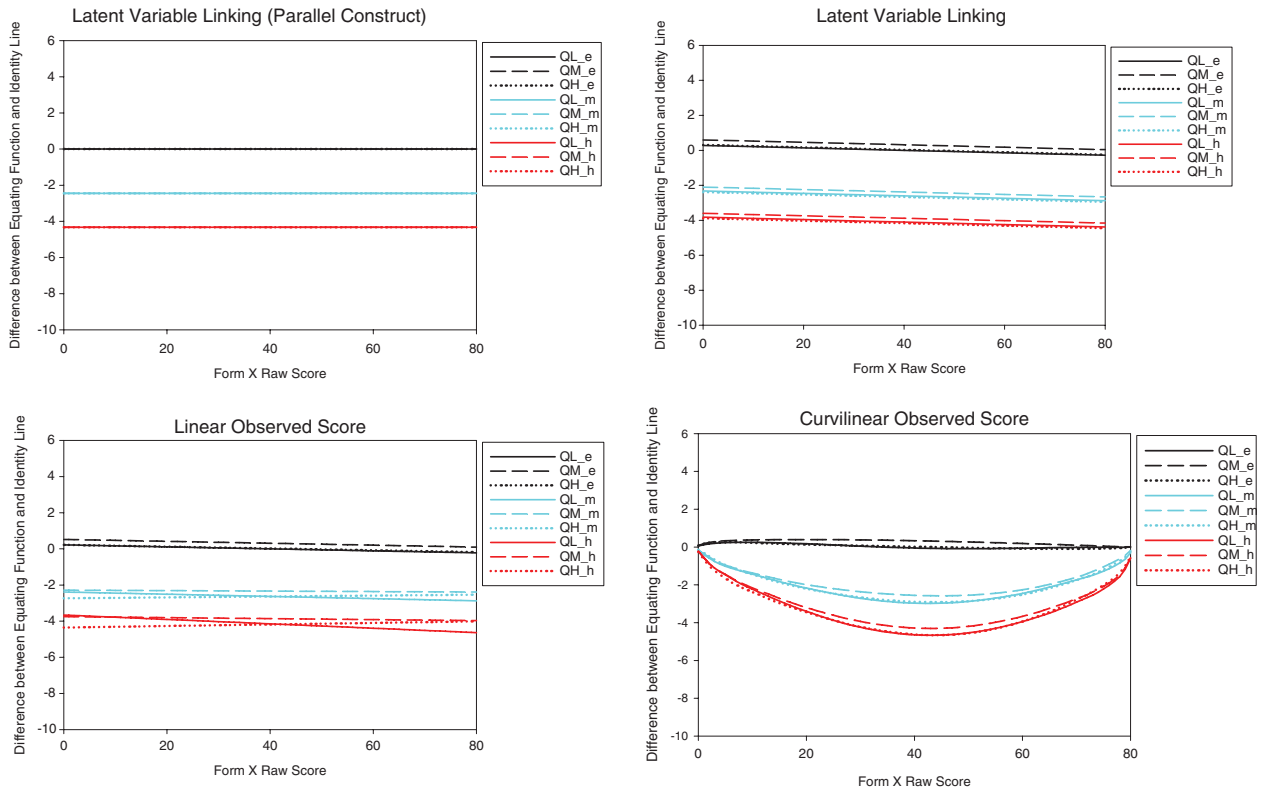

Figure 3 Linking $X_e(4{:}1)$ to $Y(3{:}2)$ when $r = 0.95$.

bowl shape of the difference between the linking functions and identity line when Form $Y(4{:}1)$ is harder than $X_e(4{:}1)$ indicates that the adjustment of the difficulty diminishes at the two ends of the score scale. In contrast to the lower left panel, which exhibits slight deviation from subpopulation invariance for the equatings of $X_e(4{:}1)$, to $Y_m(4{:}1)$ and to $Y_h(4{:}1)$, the curves in the lower right panel have small difference in the tails. This diminution of differences in the tails is a consequence of the definition of equipercentile equating, in which the relationship between the bounded observed scores is quite different from the relationship between the underlying linearly related latent true scores. In fact the linear relationship between the observed scores reflects the latent linear relationship better than the equipercentile relationship. This will be observed in subsequent figures.

Figure 2 looks much like Figure 1. The only change is that the correlation between the two latent dimensions is 0.50 instead of 0.95. The invariance with respect to correlation across Figures 1 and 2 demonstrates that equating is possible and that subpopulation invariance holds when the content structures of the tests are parallel and to a slightly lesser degree even when the tests differ in difficulty. In other words, tests that tap more than one dimension can be equated provided the content mix of items is the same across the tests, even when item difficulty might vary by a 0.15 to 0.30 standard deviations on the theta scale.

The Effects of Content Shifts on Linkings Across Subpopulations

Figure 3 depicts the difference between the linking functions of $X_e(4{:}1)$ to different versions of $Y(3{:}2)$ and the identity line in Q_L, Q_M, and Q_H, respectively, when the correlation between the two thetas is 0.95. As in Figures 1 and 2, the linking functions from the three subpopulations indicate that linking adjusts for differences in difficulty between forms. In contrast to Figures 1 and 2, the latent linear functions (upper right panel) do not exhibit constant differences across all score levels of Y. The differences lines have negative slopes. In addition, subpopulation invariance does not hold. Q_M, which differs from Q_L by 0.15 on one latent dimension and by 0.25 on the other dimension, has linking functions that are consistently higher than those obtained for Q_L and Q_H, which is 0.30 higher than Q_L on both latent dimensions. Because of this ability configuration, each version of $Y(3{:}2)$, compared to $X_e(4{:}1)$, is relatively easier for Q_M than it is for Q_L and

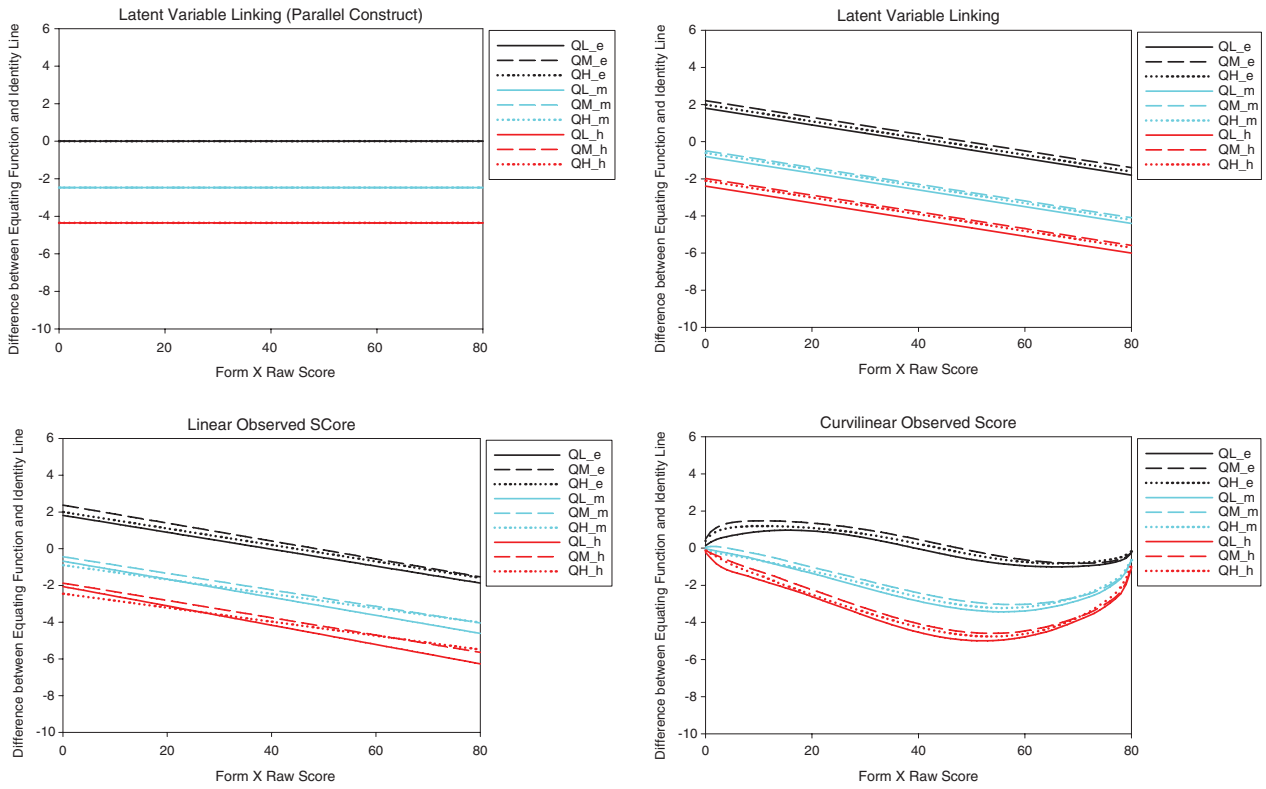

Figure 4 Linking $X_e(4{:}1)$ to $Y(3{:}2)$ when $r = 0.50$.

Q_H because content mix $Y(3{:}2)$ places greater emphasis on the ability that Q_M is relatively stronger, namely C2. Hence the relatively higher conversions in Q_M for all three levels of difficulty for content mix, $Y(3{:}2)$.

When subpopulation invariance fails to hold for scores from tests built to different specifications that are administered to the same subpopulation, linking functions can still be computed and used to link scores (Dorans & Holland, 2000). However, the linkage should be called a *concordance on a given subpopulation* rather than an *equating*, even though the calculations for the linking function are the same as those for an equating function.

The difference from the identity for the linear observed-score linkings (lower left panel) looks more similar to the differences in the linear latent score linkings than do the equipercentile linkings, as noted before. Unlike the linear latent score linking functions, however, the linear observed-score linking lines do not all have negative slopes. Once again, the slope for the Q_H linkings is positive. The equipercentile observed-score linking differences (lower right panel) are not zero even when $Y(3{:}2)$ has the same difficulty level as $X_e(4{:}1)$. As noted above the bounding of the raw score distribution produces a bowl shape.

When the correlation drops to 0.50, dramatic changes are noted in Figure 4. Relative to the baseline of parallelism (upper left panel), the linear latent score difference lines (upper right panel) deviate markedly from constant differences for all three difficulty levels, exhibiting markedly negative gradients. The linear observed-score difference lines (lower left panel) attempt to follow the linear latent score differences (upper right panel), but having less steep gradients with the Q_H relationship exhibiting a slightly smaller gradient than Q_L and Q_M. The equipercentile observed-score difference curves (lower right panel) are distorted bowls that tend, like the linear observed-score difference curves, to be higher at the low end of the score range where the curves appear to have a positive slope that turns negative in the top end of the score range by the bounded nature of equipercentile observed-score linking functions. Despite the distortion associated with the 0.50 correlation, the $Y(3{:}2)$ test remains easier relative to $X_e(4{:}1)$ in the Q_M subpopulation than in Q_L and Q_H.

Figure 5 depicts the differences in linking functions of $X_e(4{:}1)$ to the three versions of $Y(1{:}4)$ when the two fundamental latent dimensions are highly correlated. As $Y(1{:}4)$ and $X_e(4{:}1)$ have a different content mix, the linking functions of $X_e(4{:}1)$ to $Y(1{:}4)$ are not equatings. The linking functions derived in Q_L and Q_H, however, are very close to each other and noticeably lower than that observed in Q_M under all conditions. As noted in Table 4, the ability on the dimensions

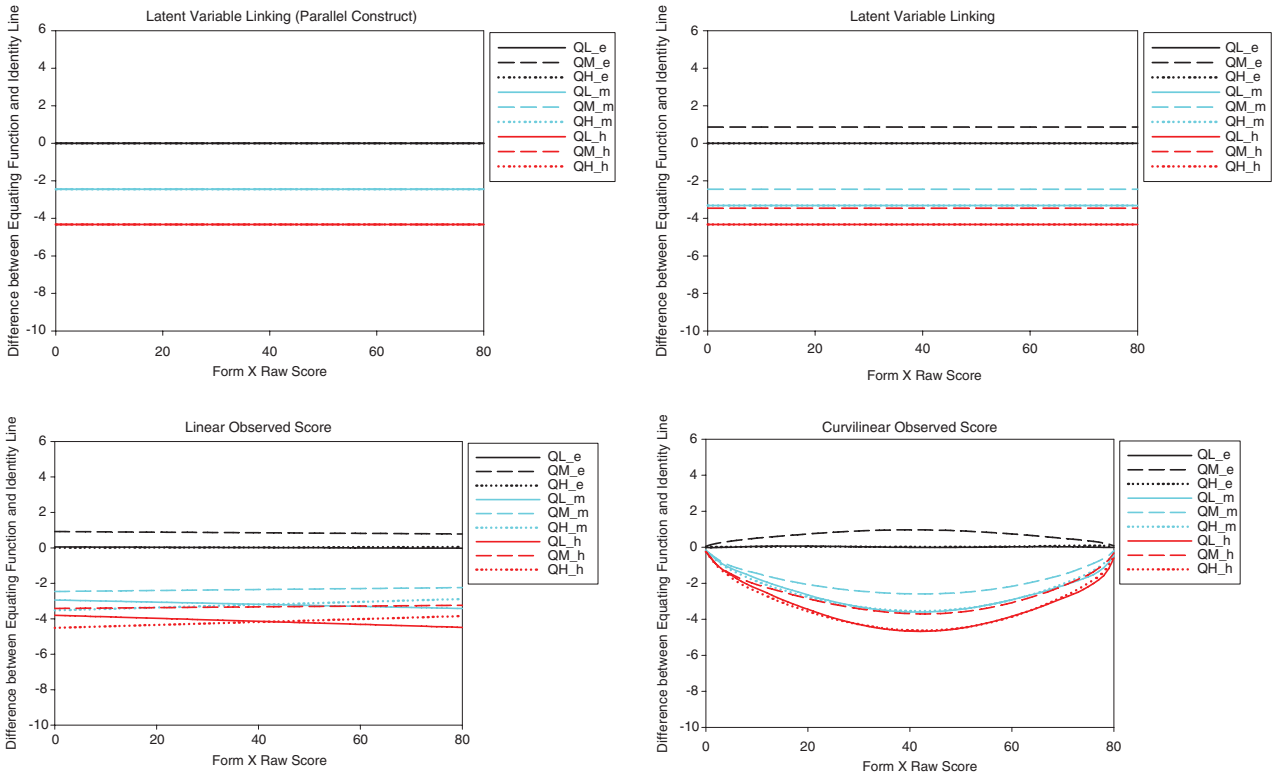

Figure 5 Linking $X_e(4{:}1)$ to $Y(1{:}4)$ when $r = 0.95$.

underlying C1 and C2 are comparable in Q_L and Q_H; in Q_M, the ability on C2 is higher than that on C1. $Y(1{:}4)$ has more items measuring C2 than $X_e(4{:}1)$ does. Q_M does relatively well on the C2 theta dimension. As a result, the three versions of the $Y(1{:}4)$ form are relatively easier in Q_M than in Q_L or Q_H. In contrast to Figures 3 and 4, the linear differences lines tend to be constant. Figure 6 shows the effect of reducing the correlation from 0.95 to 0.50. The effect is not dramatic as it was in Figure 4.

Discussion

This study examined analytically the effects of multidimensionality on latent score and observed-score linking results. The framework can be used with any number of dimensions. The special case of two dimensions was employed to understand the results of simulation studies conducted by Lin and Dorans (2011). That study used a multidimensional IRT model (Reckase, 2009) to generate simulated data and focused on developing equating functions that would be used for observed test scores. Lin and Dorans (2011) used observed-score equating methods that presume unidimensionality at the test score level. In this article, we examined equating relationships among latent test scores and how these latent linking relationships relate to observed-score linkings.

Equations 9 and 10 described the effects of correlation between underlying latent dimensions and the similarity or dissimilarity of test composition on equating functions. In the numerical examples based on Lin and Dorans (2011), we demonstrated how differences in test difficulty and content structure affects the linking relationship between two test forms that were linear combinations of the same underlying latent variables, in this case two latent variables. If the two tests had parallel structure, then the relationship between their latent total true scores was invariant across different subpopulations, even when the correlation of the latent variables was only 0.50 (see Figures 1 and 2).

As we moved away from parallel structure, the correlation mattered, as did the ability profile of the subpopulation. These effects were most evident in the case where $X_e(4{:}1)$ was linked to different versions of $Y(3{:}2)$, and invariance was not obtained in the Q_M subpopulation, which was stronger on the second latent dimension, C2, than on the first dimension, C1 (see Figures 3 and 4, where subpopulation invariance was not achieved, especially under the 0.50 correlation condition). Scores on the versions of $Y(3{:}2)$ had much less variance than $X_e(4{:}1)$, especially so when the correlation is 0.50, but even

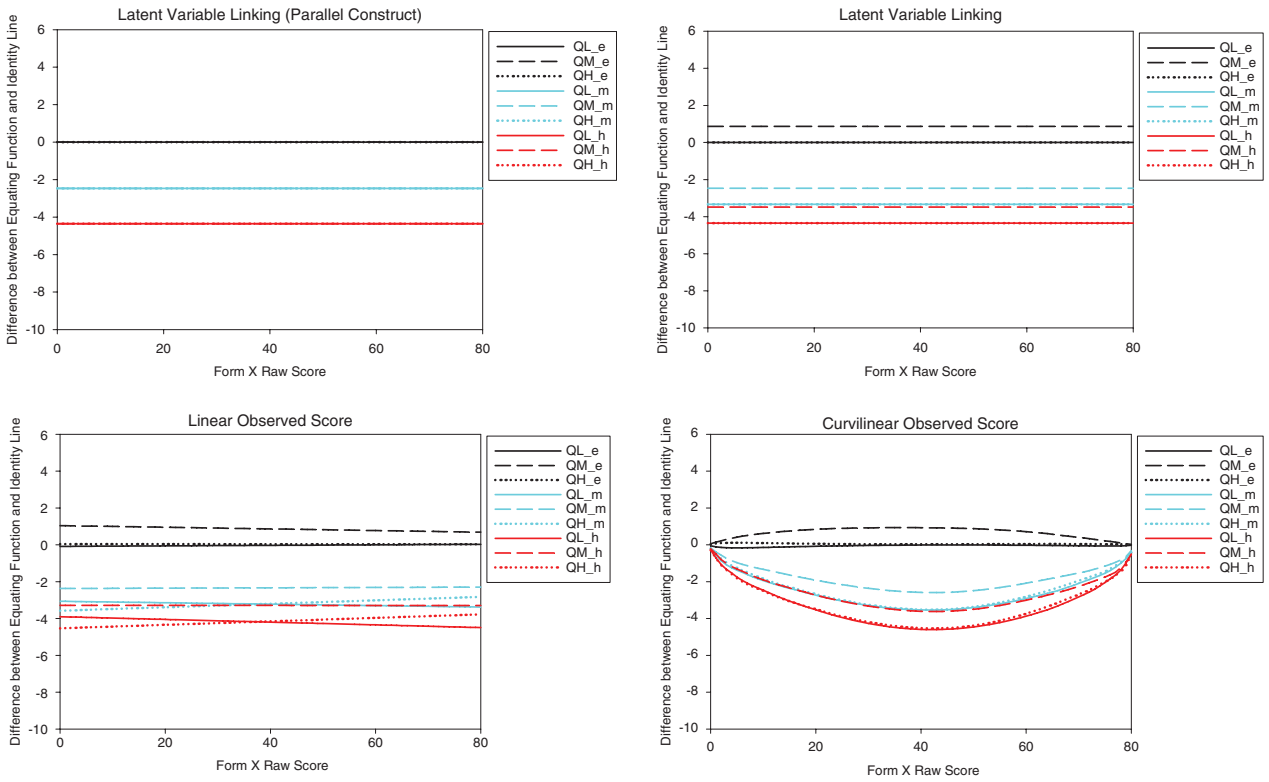

Figure 6 Linking $X_e(4{:}1)$ to $Y(1{:}4)$ when $r = 0.50$.

in the 0.95 correlation condition. Hence there was a need to compress scores on $X_e(4{:}1)$ that was reflected in the slopes in the linear panels of Figures 3 and 4 and to a more disguised manner in the curvilinear panels.

An interesting phenomenon can be noted in Figures 5 and 6, where the low ability (Q_L) and high ability subpopulations (Q_H) exhibit subpopulation invariance for tests that differed quite a bit in their structure, linking $X_e(4{:}1)$ to $Y(1{:}4)$. In Q_L and Q_H, the variances of the latent variables were 1, and the means differed by 0.30 standard deviation units on both C1 and C2 (see Table 4). In addition, the two test forms had structures that were mirror images of each other. A mirror image occurs when the proportion of items that measure different dimension flips across the tests. As a consequence, the linear combinations formed by these test forms had similar means and standard deviations in Q_L. Likewise, they had similar means and standard deviations in Q_H. Hence their linking relationships were invariant, both linear and curvilinear, across Q_L and Q_H, despite the fact that they did not correlate well with each other, especially in the 0.50 condition. However, in Q_M the linking relationship differed because the mean ability difference was 0.15 on C1 and was 0.25 on C2. This differential difference in ability accounted for the difference in elevation of the plots for this subpopulation relative to subpopulations Q_L and Q_H. Weeks (2013) addressed the effects of restrictions on structure on multidimensional linking.

A major reason for examining linkings among the latent variables is to gain a better understanding of what happens when we link observed scores. Latent variables are frequently employed in an "as is" mode, under the presumption (as if) that they are accurate descriptions of reality. They may or may not be accurate. That is a question that requires empirical resolution.

But even when they are not accurate, they remain valuable tools for "as if" modeling, which is how they were used here, because they can illuminate. Consider the simple Equations 9 and 10. In addition to helping us understand the Lin and Dorans (2011) results, they are suggestive of other findings as well. For example, they can be used to predict that in the case where Test X measures only C1 and Test Y measures only C2, the relationship between X and Y will be invariant across subpopulations even when C1 and C2 are uncorrelated provided that the structure for Y is the mirror image of the structure for X and the means and variances on C1 and C2 track each other across subpopulations. Even when X and Y are unrelated they may yield subpopulation invariant linkings. As noted in Lin and Dorans (2011) and elsewhere,

subpopulation invariance is a necessary condition for equating but not a sufficient condition. Equations 9 and 10 can provide other useful insights into what to expect with observed-score equating.

We demonstrated that score equating is possible with factorially complex tests provided the test scores are essentially tau equivalent. The strong unidimensionality requirement associated with unidimensional IRT true-score linking may be relaxed if essential tau equivalence holds at the total score level. Holland and Hoskens (2003) demonstrated that IRT can be viewed as a special case of classical test theory. Classical test theory does not make any assumptions about item-level performance. The true score is simply the expected value of performance of a test for test takers with comparable true scores. Likewise, score equating methods that total score data make no assumptions about items. Hence, observed-score equating methods have wider applicability than unidimensional IRT equating methods.

The present research also suggests that the bounded nature of observed scores causes observed-score equipercentile equating to produce a distorted reflection of the underlying relationship between latent test scores. This distorting effect merits further examination given the widespread use of equipercentile methods.

Acknowledgments

The order of the authors is alphabetical. This version and earlier versions of the article benefitted from thoughtful comments provided by Hongwen Guo, Michael Kolen, Tim Moses, Gautam Puhan, Mark Reckase, and Rebecca Zwick. The authors are particularly appreciative of the comments and advice provided by Jonathon Weeks.

Note

1 Observed performance on a test is often bounded by zero at one end and the number of items at the other end. The range of bounded true scores falls within this range of observed scores. The unbounded true score is not constrained by these boundaries.

References

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

Dorans, N. J., & Lawrence, I. M. (1987). *The internal construct validity of the SAT* (Research Report No. RR-87-35). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00239.x

Dorans, N. J., & Lawrence, I. L. (1999). *The role of the unit of analysis in dimensionality assessment* (Research Report No. RR-99-14). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1999.tb01812.x

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187 – 220). Westport, CT: American Council on Education/Prager.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, *68*, 123 – 149.

Lin, P., & Dorans, N. J. (2011, April). *Assessing population invariance of vertical linking functions*. Paper presented at National Council on Education Measurement, New Orleans, LA.

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

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*, 3 – 31.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York, NY: McGraw-Hill.

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401 – 412.

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

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago, IL: University of Chicago Press.

Weeks, J. (2013, April). Issues in multidimensional test linking. In *Linking scores in the presence of violations of unidimensionality*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Appendix

Parameter Values for the Test Forms

Table A1 Parameter Values for the Test Form X_e(4:1) and Y_e(4:1)

Item number	a_{i1}	a_{i2}	d_i
1 – 4, 41 – 44	1	0	1.75
5 – 12, 45 – 52	1	0	1
13 – 20, 53 – 60	1	0	0
21 – 28, 61 – 68	1	0	−1
29 – 32, 69 – 72	1	0	−1.75
33, 73	0	1	1.75
34 – 35, 74 – 75	0	1	1
36 – 37, 76 – 77	0	1	0
38 – 39, 78 – 79	0	1	−1
40, 80	0	1	−1.75

Table A2 Parameter Values for the Test Form Y_e(1:4)

Item number	a_{i1}	a_{i2}	d_i
1, 42	1	0	1.75
2 – 3, 43 – 44	1	0	1
4 – 5, 45 – 46	1	0	0
6 – 7, 46 – 47	1	0	−1
8, 48	1	0	−1.75
9 – 12, 49 – 52	0	1	1.75
13 – 20, 53 – 60	0	1	1
21 – 28, 61 – 68	0	1	0
29 – 36, 69 – 76	0	1	−1
37 – 40, 77 – 80	0	1	−1.75

Table A3 Parameter Values for the Test Form Y_e(3:2)

Item number	a_{i1}	a_{i2}	d_i
1 – 3, 41 – 43	1	0	1.75
4 – 9, 44 – 49	1	0	1
10 – 15, 53 – 60	1	0	0
16 – 21, 56 – 61	1	0	−1
22 – 24, 62 – 64	1	0	−1.75
25 – 26, 54 – 66	0	1	1.75
27 – 30, 67 – 70	0	1	1
31 – 34, 71 – 74	0	1	0
35 – 38, 75 – 78	0	1	−1
39 – 40, 79 – 80	0	1	−1.75

Suggested citation:

Dorans, N. J., Lin, P., Wang, W., & Yao, L. (2014). *The invariance of latent and observed linking functions in the presence of multiple latent test-taker dimensions* (ETS Research Report No. RR-14-41). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12041

Action Editor: Rebecca Zwick

Reviewers: Gautam Puhan, Hongwen Guo, Tim Moses, and Alina von Davier