Measuring the Power of Learning.®

Research Report
ETS RR–16-36

Linking Composite Scores: Effects of Anchor Test Length and Content Representativeness

Peng Lin

Neil Dorans

Jonathan Weeks

December 2016

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

Linking Composite Scores: Effects of Anchor Test Length and Content Representativeness

Peng Lin, Neil Dorans, & Jonathan Weeks

Educational Testing Service, Princeton, NJ

The nonequivalent groups with anchor test (NEAT) design is frequently used in test score equating or linking. One important assumption of the NEAT design is that the anchor test is a miniversion of the 2 tests to be equated/linked. When the content of the 2 tests is different, it is not possible for the anchor test to be adequately representative of both tests. Lin and Dorans conducted a simulation study in 2010 to investigate the effect of content representativeness of the anchor test on linking via different linking methods when the 2 tests are nonparallel in content structure in the unique case where the groups are equivalent. The current study extends the Lin and Dorans study to the case with nonequivalent group data. Specifically, the current study investigates the impact of content representativeness and length of anchor test on linking when the 2 tests are multidimensional and nonparallel in content structure. The NEAT design was employed. The linking results from 3 classic linear equating methods — Levine observed score, Tucker equating, and chained linear — were examined. The results from the study indicated that equating the tests with different structure should be avoided. For equatings with anchor test, additional bias is likely to be introduced by using an inadequate anchor test.

Keywords Nonequivalent groups with anchor test (NEAT) design; test score equating/linking; linear equating methods; Levine observed score; Tucker equating; chained linear

doi:10.1002/ets2.12122

The nonequivalent groups with anchor test (NEAT) design is frequently used in test score equating/linking. With this design, one test is administered to a group of examinees and a second test is administered to another group of examinees. Both groups take an anchor test. Ideally, the anchor test is a miniversion of the two tests to be equated/linked (Dorans, Kubiak, & Melican, 1998; Kolen & Brennan, 2004). In practice, it is often necessary to link two tests that are not parallel in content. For example, in vertical scaling, the test at each grade is developed to accommodate the curriculum emphasis of the grade, which might change across grades. As such, when the content of the two tests is different, it is not possible for the anchor test to be adequately representative of both tests (Kolen, 2007). A variety of studies have addressed the effect of the representativeness of anchor test on equating when the two tests are parallel in content structure (Cook & Petersen, 1986; Harris, 1991; Klein & Jarjoura, 1985; Sinharay & Holland, 2007; Sykes, Hou, Hanson, & Wang, 2002; Yang, 1997). However, only a few studies have investigated the effect when the two tests are not parallel in content structure (Kromrey, Parshall, & Yi, 1998; Lin, 2008; Lin & Dorans, 2010).

Lin and Dorans (2010) conducted a simulation study to investigate the effect of content representativeness of the anchor test on linking via different linking methods when the two tests are nonparallel in content structure in the unique case where the groups are equivalent. Two types of anchors were examined: an anchor that represents the reference test and an anchor that represents a proportional mix of the reference and new tests. The equating/linking methods investigated included the Tucker, Levine observe-score, chained linear, chained equipercentile, and unidimensional item response theory (IRT) true score methods. The results showed that when the two tests are nonparallel in content structure and the groups are equivalent, neither the anchor type (representing the reference test or proportional mix of reference test and new test) nor the extent of multidimensionality (correlation between content domains) showed evident impact on the equating results from most of the equating/linking methods. The IRT true score method was the exception; it was sensitive to both anchor types and multidimensionality. In the Lin and Dorans study, equivalent groups design was used for equating/linking. The group equivalence might explain why representativeness of the anchor does not matter for equipercentile and linear methods. Hence, the purpose of the present study is to extend the Lin and Dorans study to the case with nonequivalent group data. Specifically, this study investigated the impact of content representativeness and length of

Corresponding author: P. Lin, E-mail: PLin@ets.org

anchor test on linking when the two tests are multidimensional and nonparallel in content structure. The NEAT design was employed. The linking results from three classic linear equating methods — Levine observed score, Tucker equating, and chained linear — were examined.

Population Invariance in Linking Multidimensional Tests

In order to achieve the interchangeability of a useful equating, the tests must be measuring the same construct, the correlation between the two tests must be high, and the linkage must be invariant across important subpopulations (Dorans, 2004). To better understand the effects of unaccounted-for multidimensionality on equating/linking, Lin and Dorans (2011) investigated subpopulation invariance in a simulated scenario related to vertical scaling. Their hypothetical subpopulations can be thought of as (or defined by) the grade level of examinees, where the ability distribution of subpopulations varies across grades. The study attempted to simulate what might happen in the context of a vertical scaling when there is a shift in content structure, a change in test difficulty, and differential shifts in students' ability 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. Thus, the content structures of the tests are two-dimensional. At each grade level, proficiencies on the two dimensions were simulated to be either highly related or less related. They found that if the two tests have parallel content structures, the linking relationship between their observed scores is essentially subpopulation invariant regardless of the correlations between the underlying latent dimensions. The study also indicated that when there is content structure shift across tests to be linked, population invariance should not be assumed without further investigation about the characteristics of the tests and the populations that the linking functions would be applied to.

In a related study, Dorans, Lin, Wang, and Yao (2014) examined the linking relationships among latent test scores and the extent to which these relationships relate to observed-score linkings. One result of this study was the elucidation of equations that 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 can be used as a confirmatory model for the results obtained in the Lin and Dorans (2011) study. The equations also characterize the effect that the degree of correlation between the latent dimensions has on equatability as the structure departs from parallelism. The simulation model employed by Lin and Dorans and treated analytically in Dorans et al. provides a reference criterion based on a single group design that can be used to evaluate the various anchor test equatings examined in this study. These relationships will be discussed in more detail in the Methodology section.

Methodology

In this simulation study, the design followed the one employed in the Lin and Dorans (2011) study. It was assumed that two distinct content domains were tested in three subpopulations of varied ability: low (L), medium (M), and high (H). In each subpopulation, examinee proficiencies on the two dimensions might be highly related, such as with algebra and geometry, or less related, such as with math and reading. Each item in the tests was specified to load on only one content domain: either C1 or C2. The three subpopulations were referred to as a lower ability group (QL), a middle ability group (QM), and a higher ability group (QH), in correspondence to L, M, and H, respectively. Table 1 presents the generating means and standard deviations of the ability distributions (bivariate normal) for C1 and C2 for subpopulations QL, QM, and QH. Note that mean differences between QL and QH on both C1 and C2 are the same (0.30 standard deviation units). In contrast, the mean for QM on C1 is halfway between QL and QH (0.15 standard deviation units), but 0.25 above QL and 0.05 below QH on C2. The standard deviation of the abilities underlying performance on C1 and C2 is equal to 1 for all three subpopulations.

In addition, this study varied the correlation between C1 and C2 within each subpopulation to better examine the effects of unaccounted-for multidimensionality. The correlations of 0.5 and 0.95 were considered.

Nine simulated tests were developed by crossing three levels of test difficulty with three levels of content specifications. The item responses were generated using the one parameter logistic multidimensional IRT (1PL-MIRT) model. The three difficulty levels were easy (e), moderate (m), and hard (h). The mean difficulty parameter for the easy test was set at zero for items in each of the two dimensions. This is consistent with the mean ability of examinees in QL. The means of the difficulty parameters on the moderately difficult form for items in C1 and C2 were set at 0.15 and 0.25, respectively.

Table 1 Generating Ability Distribution for the Three Subpopulations

Subpopulation	Mean (C1, C2)	Standard deviation (C1, C2)
Subpopulation QL	(0, 0)	(1, 1)
Subpopulation QM	(0.15, 0.25)	(1, 1)
Subpopulation QH	(0.30, 0.30)	(1, 1)

Notes. QL = lower ability group; QM = middle ability group; QL = higher ability group.

Table 2 Combinations of Content Specification and Difficulty for the Nine Tests

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

This is consistent with the mean ability of QM. For the harder test, the means of the difficulty parameters for C1 and C2 items were both equal to 0.30. And this is consistent with the mean ability of QH. The content specifications differed with respect to how much emphasis was given to one of the two content domains, C1 and C2, for the test targeted to a particular subpopulation. For example, at ability level L, a math test might focus more on algebra than on geometry; at ability level M, the emphasis might be switched to geometry; and at ability level H, the emphasis might return to a more balanced mix of geometry and algebra. The content representation for each test was determined by the number of items measuring C1 and C2, respectively, reflecting the teaching emphasis at that specific ability level. For the three levels of content specification, the ratios 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 yielded nine tests (see Table 2). Each of these simulated tests (Y tests) was administered along with an X test in each of the three subpopulations: QL, QM, and QH, for which the tests in bold ($Y_e(4:1)$, $Y_m(1:4)$, and $Y_h(3:2)$, respectively) were targeted. The X test was designed to be an easy test, with a 4:1 ratio of C1 to C2 content representation. All of the tests contained 80 items.

In this study, $Y_e(4:1)$ was equated to the other eight Y tests with NEAT design. That is, the population taking $Y_e(4:1)$ is different from the population taking other Y tests. As $X_e(4:1)$ was administered along with all Y forms in each of the three subpopulations (QL, QM, and QH), items from $X_e(4:1)$ were picked to be external anchor tests for equating/linking.

Key Factors

Three factors were manipulated in this simulation study:

- The correlation between C1 and C2: .5, .95.
- The length of anchor test: 30 items, 20 items, 10 items.
- The content representativeness of the anchor items:
    - same content specification as the new test, $Y_e(4:1)$, and
    - proportional mix of content specification from the new and reference tests.

Several conditions of equating/linking were simulated in this study. Under all conditions, $Y_e(4:1)$ was the new test. In addition, for equating under anchor test design, QL always took $Y_e(4:1)$, whereas QM and QH only took the tests with appropriate difficulty level or content structure targeted for the population. There were three sets of conditions: Tables 3, 4, and 5 summarize these conditions.

Condition 1: Reference Test Has Same Structure as the New Test $Y_e(4:1)$ but With Different Difficulty

The upper portion of Table 3 depicts the design for equating an easier 80-item test (Y_e (4:1)), composed of four parts C1 (64 items) to one part C2 (16 items) and administered to a lower ability group (QL), to a middle difficulty form (Y_m (4:1)) administered to a middle ability group (QM) via three external anchor tests of different length: 30 items, 20 items, and

Table 3 Reference and New Tests Have Same Structure and Different Difficulty

Medium Difficulty Difference (e vs. m) and Medium Ability Difference (QL vs. QM)				
		QL(0, 0)	QM(0.15, 0.25)	QH(0.3, 0.3)
New test	Y_e(4:1)	X		
Reference test	Y_m(4:1)		X	
External anchor	(24, 6)	X	X	
	(16, 4)	X	X	
	(8, 2)	X	X	
Large Difficulty Difference (e vs. h) and Large Ability Difference (QL vs. QH)				
		QL(0, 0)	QM(0.15, 0.25)	QH(0.3, 0.3)
New test	Y_e(4:1)	X		
Reference test	Y_h(4:1)			X
External anchor	(24, 6)	X		X
	(16, 4)	X		X
	(8, 2)	X		X

Note. QL = lower ability group; QM = middle ability group; QL = higher ability group; e = easy; m = medium; h = hard.

Table 4 Reference and New Tests Have Different Content Structure and Same Difficulty

Different Content Structure (4:1 vs. 1:4) and Medium Ability Difference (QL vs. QM)					
			QL(0, 0)	QM(0.15, 0.25)	QH(0.3, 0.3)
New test		Ye(4:1)	X		
Reference test		Ye(1:4)		X	
External anchor	Representing new test	(24, 6)	X	X	
		(16, 4)	X	X	
		(8, 2)	X	X	
	Proportional mix of new and reference tests	(15, 15)	X	X	
		(10, 10)	X	X	
		(5, 5)	X	X	
Different Content Structure (4:1 vs. 3:2) and Large Ability Difference (QL vs. QH)					
			QL(0, 0)	QM(0.15, 0.25)	QH(0.3, 0.3)
New test		Y_e(4:1)	X		
Reference test		Y_e(3:2)			X
External anchor	Representing new test	(24, 6)	X		X
		(16, 4)	X		X
		(8, 2)	X		X
	Proportional mix of new and reference tests	(21, 9)	X		X
		(14, 6)	X		X
		(7, 3)	X		X

Note. QL = lower ability group; QM = middle ability group; QL = higher ability group.

10 items, respectively. The numbers in the brackets represent the number of items measuring C1 and C2, respectively, in the anchor. For example, (24, 6) means there were 30 items in the anchor test, with 24 items measuring C1 and 6 items measuring C2. The lower portion of the table parallels the upper portion and describes equating the same test (Y_e (4:1)) to an even harder reference test (Y_h (4:1)) that was administered to an even more able population (higher ability group [QH]). In all cases, the difficulty of the anchor test equaled that of the new test Y_e(4:1).

Condition 2: Reference Test Has Different Structure but With Same Difficulty as the New Form Ye(4:1)

Table 4 presents another set of simulated conditions. Here the assumption of construct equivalence is violated, such that Y_e(4:1), given in QL, is equated to an equally easy test, Y_e(1:4) given in QM, or Y_e(3:2) given in QH, both of which have

Table 5 Reference and New Tests Have Different Structure and Different Difficulty

Different Content Structure (4:1 vs. 1:4), Medium Difficulty Difference (e vs. m), and Medium Ability Difference (QL vs. QM)					
			QL(0, 0)	QM(0.15, 0.25)	QH(0.30, 0.30)
New test		Y_e(4:1)	X		
Reference test		Y_m(1:4)		X	
External anchor	Representing new test	(24, 6)	X	X	
		(16, 4)	X	X	
		(8, 2)	X	X	
	Proportional mix of new and Reference tests	(15, 15)	X	X	
		(10, 10)	X	X	
		(5, 5)	X	X	
Different Content Structure (4:1 vs. 3:2), Large Difficulty Difference (e vs. h) and Large Ability Difference (QL vs. QH)					
			QL(0, 0)	QM(0.15, 0.25)	QH(0.30, 0.30)
New form		Y_e(4:1)	X		
Reference form		Y_h(3:2)			X
External anchor	Representing new test	(24, 6)	X		X
		(16, 4)	X		X
		(8, 2)	X		X
	Proportional mix of new and reference tests	(21, 9)	X		X
		(14, 6)	X		X
		(7, 3)	X		X

Note. QL = lower ability group; QM = middle ability group; QL = higher ability group; e = easy; m = medium; h = hard.

a different content structure. Again, there are three levels of anchor test length: 30, 20, and 10. In addition, there are two types of anchor test composition: content representative of the new test and content proportional to the mix of C1 and C2 on the new and reference tests. For the anchors that represent new test, Y_e(4:1), the ratio of items measuring C1 and C2 is always 4:1. For the anchors that are content proportional to the mix of C1 and C2 of the new and reference tests, when reference form is Y_e(1:4), the ratio of items measuring C1 and C2 is the average of 4:1 and 1:4, which is (4 + 1)/2 for C1 and (1 + 4)/2 for C2, and equivalent to 1:1 on C1 and C2. Similarly, when the reference test is Y_e(3:2), the ratio of items measuring C1 and C2 is the average of 4:1 and 3:2, which is (4 + 3)/2 for C1 and (1 + 2)/2 for C2, and equivalent to 7:3 on C1 and C2.

Condition 3: Reference and New Tests Have Different Structure and Different Difficulty

The last set of conditions, as shown in Table 5, also violated the assumption of construct equivalence and incorporated difference in test difficulty. Here Y_e(4:1), given in QL, was equated to either a moderately difficult test, Y_m(1:4) given in QM, or a much more difficult test, Y_h(3:2) given in QH. Again, there were three levels of anchor test length: 30, 20, and 10; and two types of anchor test composition: content representative of the new test, and content proportional to the mix of C1 and C2 on the new and reference tests.

Data Generation

In this study, the probability of a simulee correctly answering item *i* was computed based on the 1PL-MIRT model:

$$P_s\left(BIS_i = 1 \,|\, \boldsymbol{\theta}_s\right) = \frac{e^{\left(\mathbf{a}_i'\boldsymbol{\theta}_s - d_i\right)}}{1 + e^{\left(a_i'\boldsymbol{\theta}_s - d_i\right)}}$$

where *s* denotes a simulee and d_s is a scalar denoting the difficulty of item *i*. $\boldsymbol{\theta}_s$, a 2-by-1 vector, denotes the simulee's ability. $\mathbf{a}_i$ is a 2-by-1 vector and denotes dimension weights for item *i*. As described earlier, each item measured only one dimension (C1 or C2); therefore, $\mathbf{a}_i$ is either (1, 0) when the items measured the dimension defined by C1 only, or (0,1) when the items measured solely C2 dimension. For the easy form, the means of *d* for C1 and C2 items were both 0. For the moderately difficult forms, 0.15 and 0.25 were added to the *d* values for C1 items and C2 items on the easy form

with comparable structure, respectively. For the harder forms, 0.30 was added to the *d* values for all of the items on the easy form with comparable structure. Under each combination of conditions, for each of the three subpopulations, item responses were generated for 100,000 simulees.

Linkings

For each equating/linking condition and each of the key factors, including the correlation between the latent dimensions underlying C1 and C2, linkings were conducted using three observed-score methods. This study focuses on the Levine observed score, Tucker, and chained linear method; see Kolen and Brennan (2004) for more information on these methods. In the Tucker and Levine equating, the synthetic populations were formed by weighting the new Test Sample 1 and the reference Test Sample 0. As noted before, each of these simulated tests (Y tests) were administered in each of the three subpopulations: QL, QM, and QH. Therefore, single group linking results are available for all conditions investigated in this study. As noted in Dorans et al. (2014), the single group equatings between two tests that measure different content structures are population dependent. In this study, there were two different single group linear linkings (set mean and SD equal) that served as criteria. One was from the subpopulation that took a new test in the anchor test design; the other was from the subpopulation that took the reference test. The linking results under the anchor test design, from the three equating – linking methods, are compared to the single group results. Differences from these single group equating results are reported for each of the equating/linking methods to identify the relative effect of content representativeness and the length of the anchor test for equating/linking tests under the anchor test design.

Results

A graphical representation of the results is presented in Figures A1 through A18 in the Appendix.

Table 6 shows the means and standard deviations for observed scores on the total tests and anchor tests for each of the conditions for the equatings – linkings where QL took the new test $Y_e(4{:}1)$ and QM took reference tests. Table 7 shows the same set of statistics where QL took new test $Y_e(4{:}1)$ and QH took reference tests. The tables also present the correlation between scores on the total and anchor tests. In each table, the first column indicates equating conditions: same structure, different difficulty; different structure, same difficulty; or different structure, different difficulty. The second column is the correlation between C1 and C2: 0.5 or 0.95. The next column, anchor length, depicts the length and content representativeness (C1 and C2 are the number of anchor items measuring C1 and C2, respectively) of the anchor tests. The next column presents the mean and *SD* of total and anchor scores, as well as the correlation between the two scores, for QL taking the new test $Y_e(4{:}1)$, and QM or QH taking the reference tests, under different combination of conditions. As can be seen from Table 6 and Table 7, mean of total scores is not affected by the correlation between C1 and C2, whereas *SD* of the total scores is positively affected by the correlation, that is, the *SD* of scores when correlation is 0.95 is higher than those when the correlation is 0.5, other things being equal. Another finding is that when the new and reference tests have same structure, the correlation between total and anchor tests is comparable in the new and reference tests. However, when the structures are different for new and reference tests, the correlation between total and anchor in each test depends on the representativeness of anchor. If the anchor represents the new test, the correlation between total and anchor in the new test is higher than that in the reference test; if the anchor is a content proportional to the mix of C1 and C2 on the new and reference tests, the correlation between the total and anchor in the new and reference tests is comparable. In addition, it was also found that the correlation between total and anchor is positively affected by the anchor length.

Same Structure: Equating Tests of Different Difficulty Given in Different Subpopulations

Figures A1 through A6 describe the results that occur under the same structure and different difficulty conditions, where $Y_e(4{:}1)$ administered to subpopulation QL is equated to $Y_m(4{:}1)$ (moderate difference in difficulty) administered to subpopulation QM, and $Y_h(4{:}1)$ (large difference in difficulty) administered to subpopulation QH, respectively. The left panels in Figures A1, A3, and A5 show the difference between the single group criterion equatings in QL and QM when $Y_e(4{:}1)$ is equated to $Y_m(4{:}1)$. The difference is fairly close to zero in both the cases where the two latent variables, C1 and C2, correlate .5 and .95. The positive slope of the difference indicates that the slope of the single group equating relationship in the QM is slightly higher than the slope in the QL. Though barely discernible, there is a slight shift in the intercept of

Table 6 Descriptive Statistics for Equating Where a Lower Ability Group (QL) Take New Test Y_e(4:1) and a Middle Ability Group (QM) Take Reference Tests

Equating Condition	Corr (C1, C2)	Anchor Length			QL					QM				
					Total		Anchor		Corr (Total, Anchor)	Total		Anchor		Corr (Total, Anchor)
		Tot	C1	C2	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Same Structure, Different Difficulty Ye(4:1) to Ym(4:1)	0.5	30	24	6	39.99	13.26	15.00	5.36	0.87	39.94	13.34	15.88	5.36	0.87
		20	16	4			9.99	3.75	0.83			10.60	3.75	0.83
		10	8	2			5.00	2.07	0.74			5.29	2.08	0.74
	0.95	30	24	6	40.04	14.33	15.01	5.75	0.88	40.01	14.32	15.91	5.74	0.88
		20	16	4			10.01	4.01	0.85			10.61	3.98	0.85
		10	8	2			5.01	2.19	0.77			5.30	2.17	0.77
Different Structure, Same Difficulty Ye(4:1) to Ye(1:4)	0.5	30	24	6	39.99	13.26	15.00	5.36	0..87	43.15	13.27	15.88	5.36	0.69
		20	16	4			9.99	3.75	0.83			10.60	3.75	0.66
		10	8	2			5.00	2.07	0.74			5.29	2.08	0.58
		30	15	15			14.99	5.25	0.86			16.02	5.01	0.81
		20	10	10	39.99	13.26	9.99	3.76	0.82	43.15	13.27	10.69	3.56	0.77
		10	5	5			5.00	1.96	0.72			5.33	1.93	0.68
	0.95	30	24	6	40.04	14.33	15.01	5.75	0.88	43.23	14.23	15.91	5.74	0.87
		20	16	4			10.01	4.01	0.85			10.61	3.98	0.84
		10	8	2			5.01	2.19	0.77			5.30	2.17	0.75
		30	15	15			15.02	5.74	0.88			16.04	5.55	0.88
		20	10	10	40.04	14.33	10.01	4.10	0.85	43.23	14.23	10.70	3.91	0.85
		10	5	5			5.00	2.09	0.76			5.33	2.07	0.75
Different Structure, Different Difficulty Ye(4:1) to Ym(1:4)	0.5	30	24	6	39.99	13.26	15.00	5.36	0.87	39.98	13.33	15.88	5.36	0.69
		20	16	4			9.99	3.75	0.83			10.60	3.75	0.66
		10	8	2			5.00	2.07	0.74			5.29	2.08	0.59
		30	15	15			14.99	5.25	0.86			16.02	5.01	0.81
		20	10	10	39.99	13.26	9.99	3.76	0.82	39.98	13.33	10.69	3.56	0.77
		10	5	5			5.00	1.96	0.72			5.33	1.93	0.68
	0.95	30	24	6	40.04	14.33	15.01	5.75	0.88	40.03	14.31	15.91	5.74	0.87
		20	16	4			10.01	4.01	0.85			10.61	3.98	0.84
		10	8	2			5.01	2.19	0.77			5.30	2.17	0.75
		30	15	15			15.02	5.74	0.88			16.04	5.55	0.88
		20	10	10	40.04	14.33	10.01	4.10	0.85	40.03	14.31	10.70	3.91	0.85
		10	5	5			5.00	2.09	0.76			5.33	2.07	0.75

single group equating in QL and QM as well, with the difference between the two equatings intersecting the baseline of zero difference at 40 in the $R_{C1C2} = .95$ condition and slightly higher than 40 in the $R_{C1C2} = .5$ condition. The left panels in Figures A2, A4, and A6 show the difference between the single group criterion equatings in QL and QH when Y_e(4:1) is equated to Y_h(4:1), where difference in test difficulty is larger, and its effect on the lack of population invariance of the single group criterion is quite noticeable. Lin and Dorans (2011) found that in QL, the harder reference test has a lower SD than the easier new test, which was designed for QL. In QH, the easier new test has a lower SD than the harder reference test, which was designed for QH. These differences account for the positive slopes in the left panels. Note again that the $R_{C1C2} = .95$ case yields a zero difference closer to 40 than the $R_{C1C2} = .50$ case.

The results obtained using the Levine observed-score method are depicted in Figure A1 (moderate difference in difficulty) and Figure A2 (large difference in difficulty). In Figure A1, the difference in criterion slopes is reflected in the panels of anchor test equating (middle and right panels in Figure A1) that use the QL and QM single group equatings as criteria, respectively. The middle panels in Figure A1 indicate that the anchor test equatings have higher slopes relative to the QL single group criterion; the slopes of the anchor test equatings in the right-hand panels are less steep than the middle panels when compared to the QM single group criterion. For the $R_{C1C2} = .5$ condition, the length of the anchor has little or no effect on anchor test equatings. However, for the $R_{C1C2} = .95$ condition, anchor test equating with longer anchors yielded results closer to the single group criteria. In Figure A2, the middle and right panels tell similar versions of the same story, with the only appreciable difference being the tilt of the single group criterion. Setting these differences aside, it is clear that in all but the case of the 10-item anchor in the $R_{C1C2} = .95$ correlation condition, the number of anchor items has

Table 7 Descriptive Statistics for Equatings Where a Lower Ability Group (QL) Take New Test Y_e(4:1) and a Higher Ability Group (QH) Take Reference Tests

Equating Condition	Corr (C1, C2)	Anchor Length			QL					QH				
					Total		Anchor		Corr (Total, Anchor)	Total		Anchor		Corr (Total, Anchor)
		Tot	C1	C2	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Same Structure, Different Difficulty Ye(4:1) to Yh(4:1)	0.5	30	24	6	39.99	13.26	15.00	5.36	0.87	40.08	13.27	16.61	5.32	0.87
		20	16	4			9.99	3.75	0.83			11.08	3.71	0.83
		10	8	2			5.00	2.07	0.74			5.52	2.06	0.74
	0.95	30	24	6	40.04	14.33	15.01	5.75	0.88	40.02	14.29	16.57	5.70	0.88
		20	16	4			10.01	4.01	0.85			11.05	3.97	0.85
		10	8	2			5.01	2.19	0.77			5.52	2.16	0.77
Different Structure, Same Difficulty Ye(4:1) to Ye(3:2)	0.5	30	24	6	39.99	13.26	15.00	5.36	0..87	40.22	12.58	16.61	5.32	0.84
		20	16	4			9.99	3.75	0.83			11.08	3.71	0.81
		10	8	2			5.00	2.07	0.74			5.52	2.06	0.72
		30	21	9			14.99	5.25	0.86			16.61	5.20	0.85
		20	14	6	39.99	13.26	9.99	3.76	0.82	44.22	12.58	11.11	3.72	0.82
		10	7	3			5.00	1.96	0.72			5.50	1.94	0.71
	0.95	30	24	6	40.04	14.33	15.01	5.75	0.88	44.17	14.13	16.57	5.70	0.88
		20	16	4			10.01	4.01	0.85			11.05	3.97	0.85
		10	8	2			5.01	2.19	0.77			5.52	2.16	0.76
		30	21	9			15.02	5.74	0.88			16.57	5.69	0.88
		20	14	6	40.04	14.33	10.01	4.10	0.85	44.17	14.13	11.08	4.05	0.85
		10	7	3			5.00	2.09	0.76			5.49	2.07	0.76
Different Structure, Different Difficulty Ye(4:1) to Yh(3:2)	0.5	30	24	6	39.99	13.26	15.00	5.36	0.87	40.07	12.67	16.61	5.32	0.84
		20	16	4			9.99	3.75	0.83			11.08	3.71	0.81
		10	8	2			5.00	2.07	0.74			5.52	2.06	0.72
		30	21	9			14.99	5.25	0.86			16.61	5.20	0.85
		20	14	6	39.99	13.26	9.99	3.76	0.82	40.07	12.67	11.11	3.72	0.82
		10	7	3			5.00	1.96	0.72			5.50	1.94	0.71
	0.95	30	24	6	40.04	14.33	15.01	5.75	0.88	40.00	14.22	16.57	5.70	0.88
		20	16	4			10.01	4.01	0.85			11.05	3.97	0.85
		10	8	2			5.01	2.19	0.77			5.52	2.16	0.76
		30	21	9			15.02	5.74	0.88			16.57	5.69	0.88
		20	14	6	40.04	14.33	10.01	4.10	0.85	40.00	14.22	11.08	4.05	0.85
		10	7	3			5.00	2.09	0.76			5.49	2.07	0.76

only a slight effect on the results. These results are consistent with the assumption of a common true score that underlies the Levine equating model. This assumption, which implies a perfect true score correlation between the anchor score and each total test score, seems to compensate for differences in the observed correlations between anchor and total scores due to differences in the length of the anchor.

Figures A3 and A4 include results for the Tucker equating method. As expected, the length of the anchor matters quite a bit, as does the difference in test difficulty and subpopulation ability. The correlations between the anchor and total are .87/.88 for the 30-item anchor, .83/.85 for the 20-item anchor, and .74/.77 for the 10-item anchor. (The correlation before the "/" sign is for the $R_{C1C2} = .5$ case, and the correlation after it is for the $R_{C1C2} = .95$ case.) The longer anchor, the less difference between the anchor test equating and the single group criteria. Further, the difference is in the expected direction. The Tucker method assumes that QL and QM or QH are equally able unless the anchor information indicates otherwise. The Tucker method discounts the group ability difference indicated by anchor scores to the degree that the anchor total correlation deviates from unity. As such, the tilt seen in Figure A2 is also evident in Figure A4, where the effects of a short anchor on difference between anchor test equating and single group criteria are more pronounced than in Figure A3 due to the larger difficulty/ability difference between the tests/samples.

Figures A5 and A6 include the results for the chained linear method. Typically, the chained results fall between the Levine and Tucker results. The chained method takes the anchor at face value regardless of the anchor/total correlation. The Levine method, on the other hand, attributes any deviation from a perfect correlation to the fact that the anchor is unreliable and makes an adjustment accordingly, in essence treating anchor score differences as underestimates of true

differences. As noted above, the Tucker method discounts the anchor score difference to the degree that its correlation with the total test scores deviates from unity. Hence the chained method often falls between the Tucker and Levine methods. In this case, however, the effect of the anchor under the chained approach appears to be an attenuated version of the Tucker method. It too exhibits a dependency on the length of the anchor, which is not evident in the Levine method.

Different Structure and Same Difficulty

Figures A7 through A12 describe the results that occur under the different structure and same difficulty conditions, where $Y_e(4{:}1)$ is administered to subpopulation QL, $Y_e(1{:}4)$ to QM, and $Y_e(3{:}2)$ to QH, respectively. The left panels in Figures A7, A9, and A11 show the difference between the single group criterion equatings in QL and QM when $Y_e(4{:}1)$ is equated to $Y_e(1{:}4)$. The difference is fairly large compared to those observed in same structure, different difficulty conditions, in both cases where the two latent variables, C1 and C2, correlate 0.5 and 0.95. The negative slope of the difference indicates that the slope of the single group equating relationship in the QM is slightly lower than the slope in the QL. The positive difference between the single group criterion equatings from QL and QM through the whole score range indicates that, compared with the reference test $Y_e(1{:}4)$, new test $Y_e(4{:}1)$ is more difficult for QM than for QL. The differential ability difference on C1 and C2 between QM and QL could explain this difference. As noted before, the means of ability on C1 and C2 are both 0 in QL, and those on C1 and C2 are both 0.30 standard deviation units higher in QH. In contrast, the mean on C1 in QM is 0.15 standard deviation units higher than QL, but 0.25 above QL on C2. QM is more able on C2 than C1, compared with the QL and QH. As the reference test $Y_e(1{:}4)$ has more items measuring C2, it is relatively easier for QM than for QL and QH compared with $Y_e(4{:}1)$, which has more items measuring C1. The left panels in Figures A8, A10, and A12 show the difference between the single group criterion equatings in QL and QH when $Y_e(4{:}1)$ is equated to $Y_e(3{:}2)$. The difference is small, especially when $R_{C1C2} = .95$. In addition, the slope of the difference is close to 0 in both cases, $R_{C1C2} = 0.5$ and 0.95. This might be because the content structures of the two tests are closer and the ability differences on C1 and C2 in QL and QH are comparable.

Figures A7 and A8 describe the results for the Levine method. Figure A7, where $Y_e(4{:}1)$ to QL is equated to $Y_e(1{:}4)$ to QM, indicates that content representativeness of anchor makes a difference for the equating results for the Levine method. There is a separation between the anchors that is representative of the new test, multiples of (8:2), versus those that are a proportion mix of the two tests, multiples of (5:5). Also, the choice of single group criterion (in QL or QM) is very important for evaluating the equating results, due to the lack of population invariance. Different conclusions about the content representativeness of anchor, anchor length, and correlation between C1 and C2 can be made based on the middle and right panels of Figure A7. Figure A8 indicates that the Levine results are close to the single group criterion equatings when $Y_e(4{:}1)$ administered to QL is linked to $Y_e(3{:}2)$ to QH; the 10-item anchor (8:2) does not perform as well as the other anchors in the $R_{C1C2} = .95$ case.

Figures A9 and A10 contain the Tucker results. In Figure A9, where $Y_e(4{:}1)$ administered to QL is equated to $Y_e(1{:}4)$ to QM, the effect of the correlation between C1 and C2 is evident, as is the effect of the anchor length. The anchor equating results are closer to the single group criterion linkings with higher correlation between C1 and C2 and, possibly, longer anchor tests. In addition, the equating results with the anchors that are a proportional mix of new and reference tests are closer to the single group criteria than those representative of the new test, other things being equal. Figure A10, where $Y_e(4{:}1)$ administered to QL is equated to $Y_e(3{:}2)$ to QH, also indicates that the anchor test equatings are affected by the correlation between C1 and C2 and anchor length. But the effect of content representativeness of anchor is only evident when $R_{C1C2} = 0.5$ and the anchor length is 20 or 10.

The chained linear results are shown in Figures A11 and A12. As noted earlier, they are essentially an attenuated version of the Tucker results.

Different Structure and Different Difficulty

Figures A13 through A18 describe the results that occur under the different structure and different difficulty conditions, where $Y_e(4{:}1)$ is administered to QL, $Y_m(1{:}4)$ to QM, and $Y_h(3{:}2)$ to QH, respectively. The left panels in Figures A13, A15, and A17 show the difference between the single group criterion equatings in QL and QM when $Y_e(4{:}1)$ is equated to $Y_m(1{:}4)$. The positive slope of the difference indicates that the slope of the single group equating relationship in the QM is slightly higher than the slope in the QL. The positive difference between the single group criterion equatings from QL

and QM through the whole score range indicates that, compared with the reference test $Y_m(1{:}4)$, the new test $Y_e(4{:}1)$ is more difficult for QM than for QL. This is consistent with what was found in different structure same difficulty conditions, possibly because the reference test $Y_m(1{:}4)$ has more items measuring C2 where QM is relatively stronger. The left panels in Figures A14, A16, and A18 show the difference between the single group criterion equatings in QL and QH when $Y_e(4{:}1)$ is equated to $Y_h(3{:}2)$. The positive slope of the difference indicates that the slope of the single group equating relationship in the QH is slightly higher than the slope in the QL. There is shift in the intercept of single group equating in QL and QH as well, with the difference between the two equatings intersecting the baseline of zero difference at 40 in the $R_{C1C2} = .95$ condition and around 30 in the $R_{C1C2} = .5$ condition.

Figures A13 and A14 describe the results for the Levine method. Figure A13, where $Y_e(4{:}1)$ administered to QL is equated to $Y_m(1{:}4)$ to QM, tells a similar story to the one in Figure A7, where $Y_e(4{:}1)$ to QL is equated to $Y_e(1{:}4)$ to QM. That is, anchor test equatings from the Levine method are affected by content representativeness and length of anchor test, as well as the correlation between C1 and C2, when structure difference between the new and reference tests is large (4:1 vs. 1:4) and the subpopulations have differential ability difference on C1 and C2 ([0,0] vs. [0.15, 0.25]). However, the evaluation of the effects depends on the choice of single group criterion. For example, if single group equating in QL is used as the criterion, the equatings with anchors that are a proportional mix of new and reference tests are better than those with anchors that represent a new test. However, if the single group equating in QM is used as the criterion, the opposite conclusion would be made. Figure A14, where $Y_e(4{:}1)$ to QL is equated to $Y_h(3{:}2)$ to QH, also tells a story similar to the one in Figure A8, where $Y_e(4{:}1)$ to QL is equated to $Y_e(3{:}2)$ to QH. That is, the anchor test equatings from the Levine method exhibit little dependence on content representativeness of anchor test, anchor length, and the correlation between C1 and C2, when the structure change is not that large (4:1 vs. 3:2) and the subpopulations have common ability difference on C1 and C2 ([0,0] vs. [0.3, 0.3]) . One exception is when $R_{C1C2} = .95$ and anchor length is 10.

The differences in total anchor correlations noted above may account for the differences observed in Figures A15 and A16, and to a lesser extent in Figures A17 and A18, where the dependencies on anchor length and anchor representativeness are greater for both the Tucker and chained methods in the $R_{C1C2} = .5$ case, compared to the $R_{C1C2} = .95$ case. In addition, the effects of content representativeness of anchor tests are more evident in the $Y_e(4{:}1)$ to $Y_m(1{:}4)$ anchor test equatings than those in the $Y_e(4{:}1)$ to $Y_h(3{:}2)$ equatings. The findings are similar to those shown in Figures A9 to A12.

Discussion

In every case, the Levine method tracks the results of the single group criterion better than the Tucker or chained methods and, with the exception of the 10-item anchor, tends to be the least affected by the length and representativeness of the anchor. This is because the Levine method conceptually attributes any deviation from a perfect correlation between the total and anchor score to unreliability of the anchor. The method relies heavily on the assumption that the anchor score and the total test score measure the same thing, an assumption that works well in all situations but the one where the structures are quite divergent.

The Tucker method, on the other hand, is considerably more sensitive to the anchor. In short, Tucker assumes that the two groups taking the test are equal in ability and proceeds in this manner unless information in the anchor provides evidence to the contrary; the Tucker method discounts the anchor score to the degree that its correlation with the total test scores deviates from unity. As such, there are notable departures from the single group criterion when Tucker is used. This is particularly salient in cases where the anchor is very short.

As noted earlier, the chained method tends to fall between the Levine and Tucker results. This is because the approach does not overemphasize or underemphasize the information in the anchor. However, in terms of the pattern of differences between the chained and criterion results, the chained method tracks more closely with the Tucker method than the Levine method, with the key difference being that the effect of the anchor length is more attenuated.

The effect of using a proportional mix for the anchor or a mix representative of the new test seems more nuanced. With some methods, in certain cases the proportional mix anchor tracks more closely with the criterion, while for other methods and cases the representative-of-new-test anchor tracks better, depending on the structure difference and choice of single group criterion.

When considering the interaction between the various design conditions and the equating methods, it is important to consider the results from the single group equating. These results provide a useful reference for understanding the effect of differences in content structure and difficulty without any consideration of an anchor. When the tests measure

the same construct the group invariance of single group equating holds well, although there is some difference in the cases where the latent variables are correlated at .5. The difference is quite obvious in the conditions where the structure is different, and more so in the cases where the construct representation shifts from (4:1) to (1:4), relative to the shift to (3:2). Again, the difference is more pronounced when the latent variables are correlated at .5. These results alone are sufficient to suggest that equating the tests with different structure should be avoided. The corresponding results in the anchor test linkings – equatings, relative to the single group criterion, also illustrate that additional bias is likely to be introduced by using an inadequate anchor test.

References

Cook, L., & Petersen, N. (1986). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225 – 244.

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

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating* (Statistical Report No. SR-98-02). Princeton, NJ: Educational Testing Service.

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* (Research Report No. RR-14-41). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12041

Harris, D. (1991, April). *Equating with nonrepresentative common item sets and nonequivalent groups*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Klein, L., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, *22*, 197 – 206.

Kolen, M. J. (2007). Data collection design and linking procedures. In N. Dorans & P. Holland, (Eds.), *Linking and aligning scores and scale* (pp. 31 – 55) New York, NY: Springer.

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

Kromrey, J., Parshall, C., & Yi, Q. (1998, April). *The effects of content representativeness and differential weighting on test equating: A Monte Carlo Study*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Lin, P. (2008). *IRT vs. factor analysis approaches in analyzing multigroup multidimensional binary data: The effect of structural orthogonality, and the equivalence in test structure, item difficulty, & examinee groups* (Unpublished dissertation). University of Maryland, College Park.

Lin, P., & Dorans, N. J. (2010, April). *The effect of content representativeness of anchor items on linking forms not parallel in content using NEAT design*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

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

Sinharay, S., & Holland, P. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*, 249 – 275.

Sykes, R., Hou, L., Hanson, B., & Wang, Z. (2002, April). *Multidimensionality and the equating of a mixed-format math examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Yang, W. (1997, March). *The effects of content mix and equating method on the accuracy of test equating using anchor-item design*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Appendix

Each figure contains six panels. The top three panels show the results obtained in the case where the underlying dimensions, C1 and C2, are correlated at the .5 level; the bottom three panels correspond to the case where underlying dimensions are correlated at the .95 level (a case of little multidimensionality). As mentioned previously, one important finding in Lin and Dorans (2011) was that when the content structure of the two tests to be linked is not parallel, population invariance of linking results should not be assumed without further investigation into the characteristics of the test and subpopulations. Thus, single group linking was conducted in QL, the subpopulation that took the new test Y_e(4:1), and in QM or QH, whichever subpopulation took the reference test, separately. The left column in each set of three panels illustrates the difference between the two single group criterion equatings, namely those obtained in QL and QM or QH, in particular

QM – QL or QH – QL. The next two columns represent the results of the anchor test equating using a particular method: Levine observed score, Tucker, or chained linear. In the middle column, the anchor test equatings are compared to the single group criterion equating in QL. In the right column, the anchor test equatings are compared to the single group criterion equating in QM or QH. Each line in the panels corresponds to the difference in linking functions between what was obtained by a particular anchor type (a particular combination of anchor test length and content representativeness) under anchor test design and the single group criterion linking. In the figures where tests of same content structure are linked, the content representativeness has only one level because the proportional mix of reference and new tests is equivalent to the one representing the new test only. The horizontal axis in each of the panels is the total score on the new test, $Y_e(4:1)$, and the vertical axis is the difference between the linking functions.

There are 18 figures. The first six figures (A1 through A6) depict the situation in which equating is permissible, that is, when the content structure between the tests is the same but difficulty differs (i.e., the conditions in Table 3). There are two figures for each of the three linear anchor test equating methods. One figure is for equating $Y_e(4:1)$ to $Y_m(4:1)$, and the other one is for equating $Y_e(4:1)$ to $Y_h(4:1)$. The next six figures (A7 through A12) contain the results that occur when tests have different content structure and same difficulty (i.e., the conditions in Table 4). Again, there are two figures for each of the three linear anchor test equating methods. The last six figures (A13 throu gh A18) contain the results from linking tests of different content structure and different difficulty (i.e., the conditions in Table 5). There are two figures for each of the three linear equating methods.

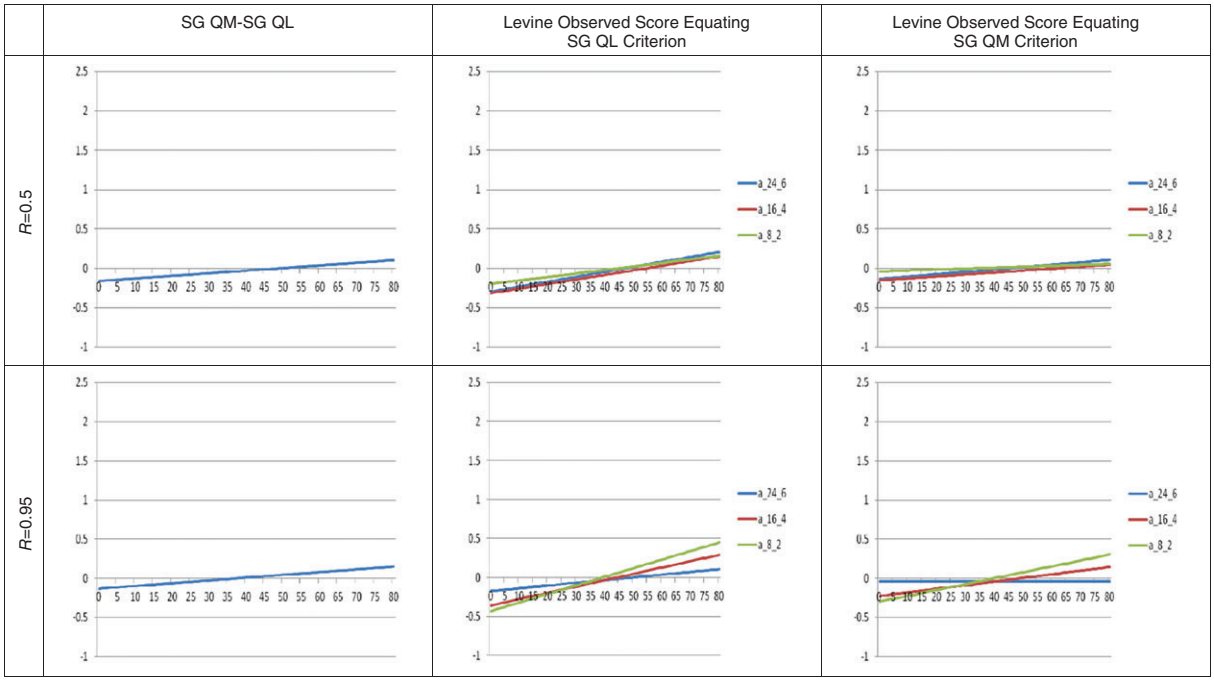

Figure A1 Same structure different difficulty $Y_e(4:1) \geq Y_m(4:1)$. Differential ability difference QL QM. Levine observed-score equating.

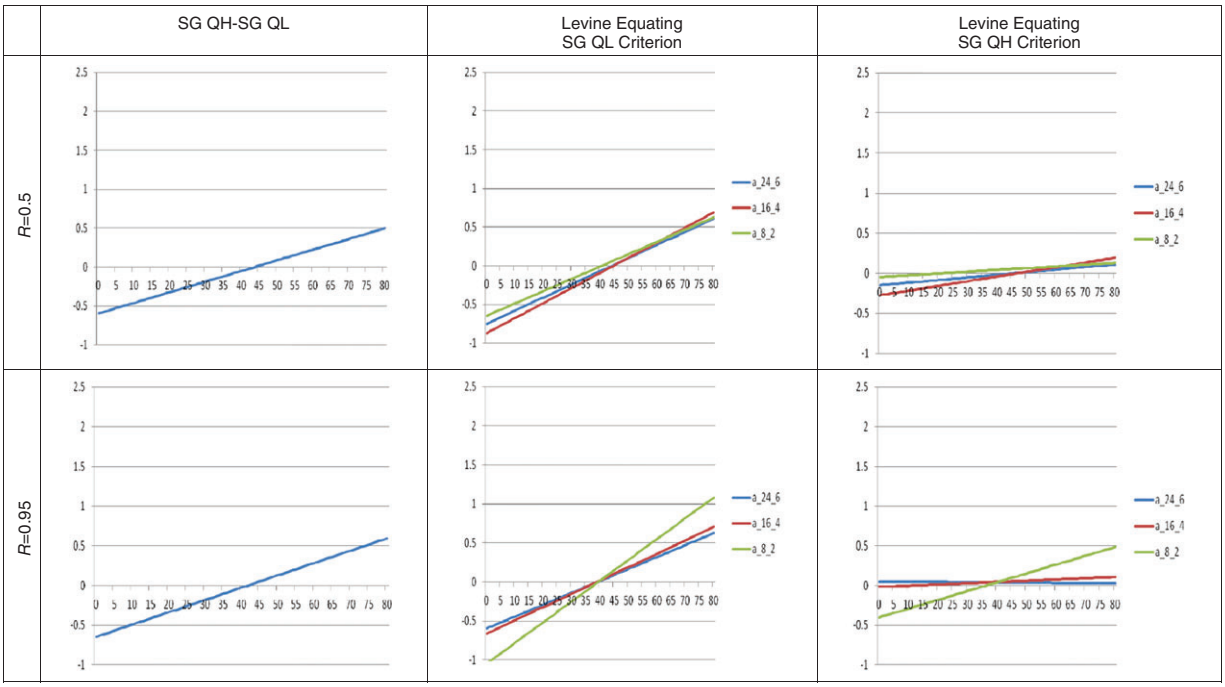

Figure A2 Same structure different difficulty $Y_e(4{:}1) \geq Y_h(4{:}1)$. Common ability difference QL QH. Levine observed-score equating.

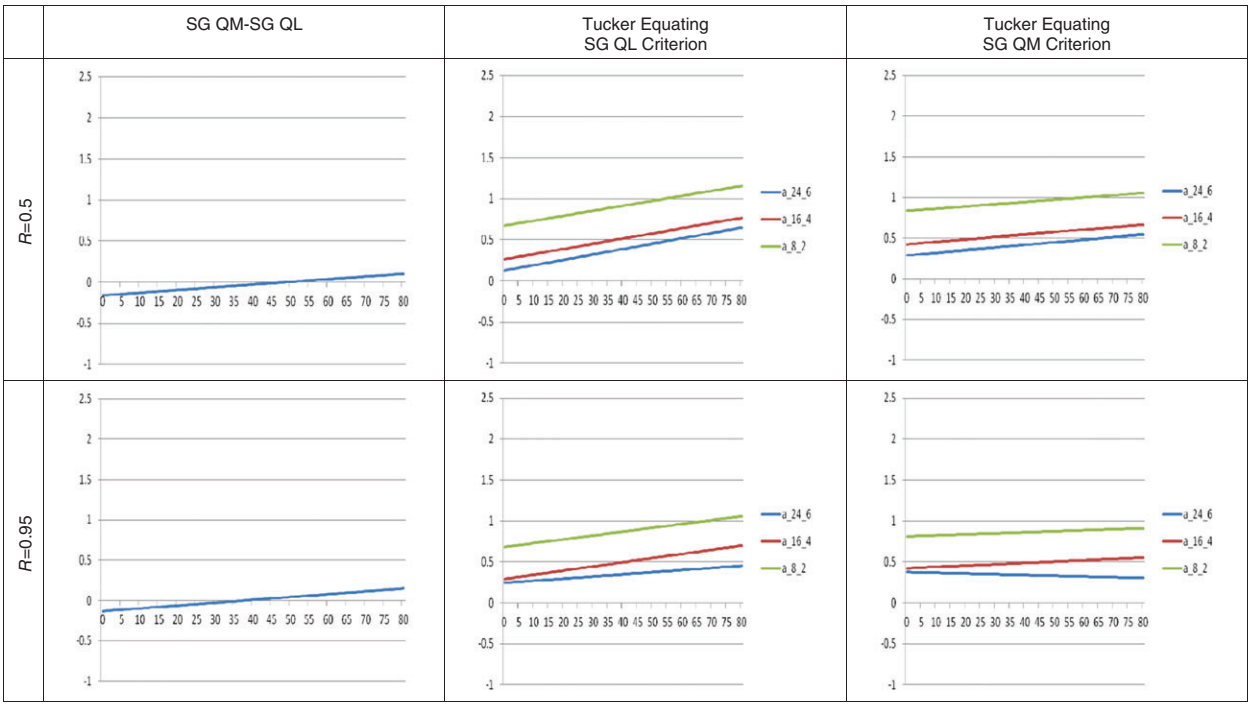

Figure A3 Same structure different difficulty $Y_e(4{:}1) \geq Y_m(4{:}1)$. Differential ability difference QL QM. Tucker equating.

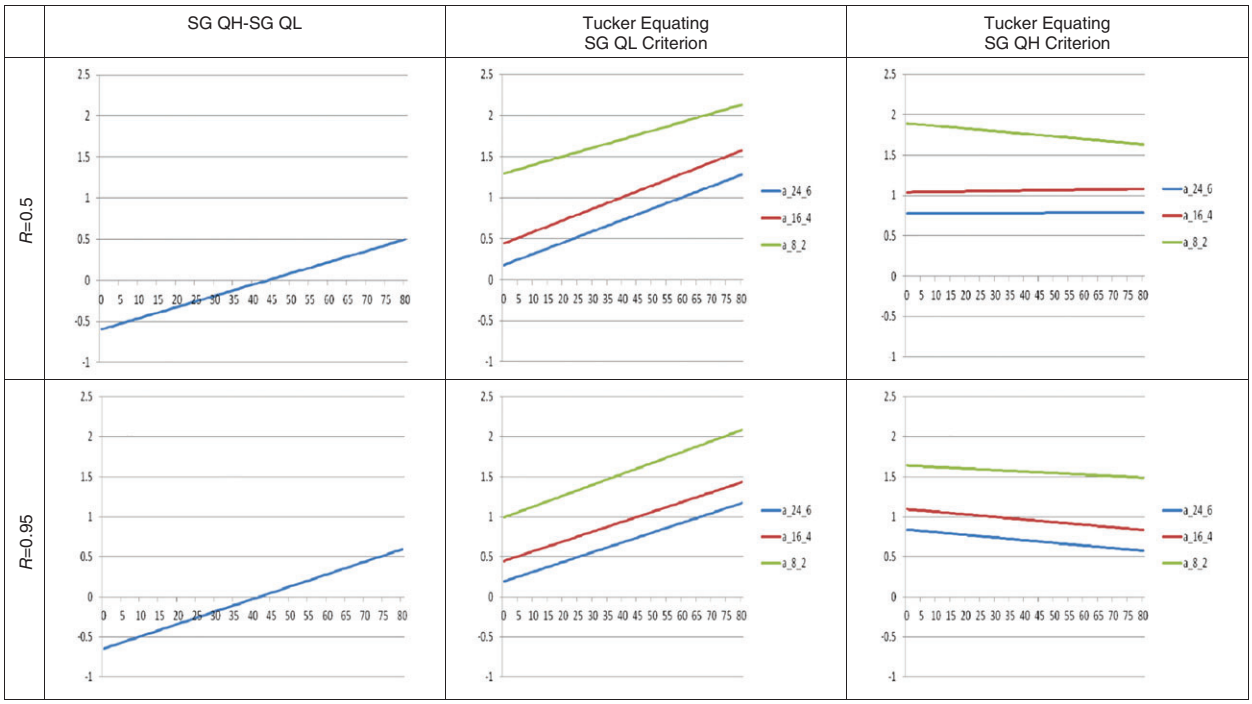

Figure A4 Same structure different difficulty $Y_e(4{:}1) \geq Y_h(4{:}1)$. Common ability difference QL QH. Tucker equating.

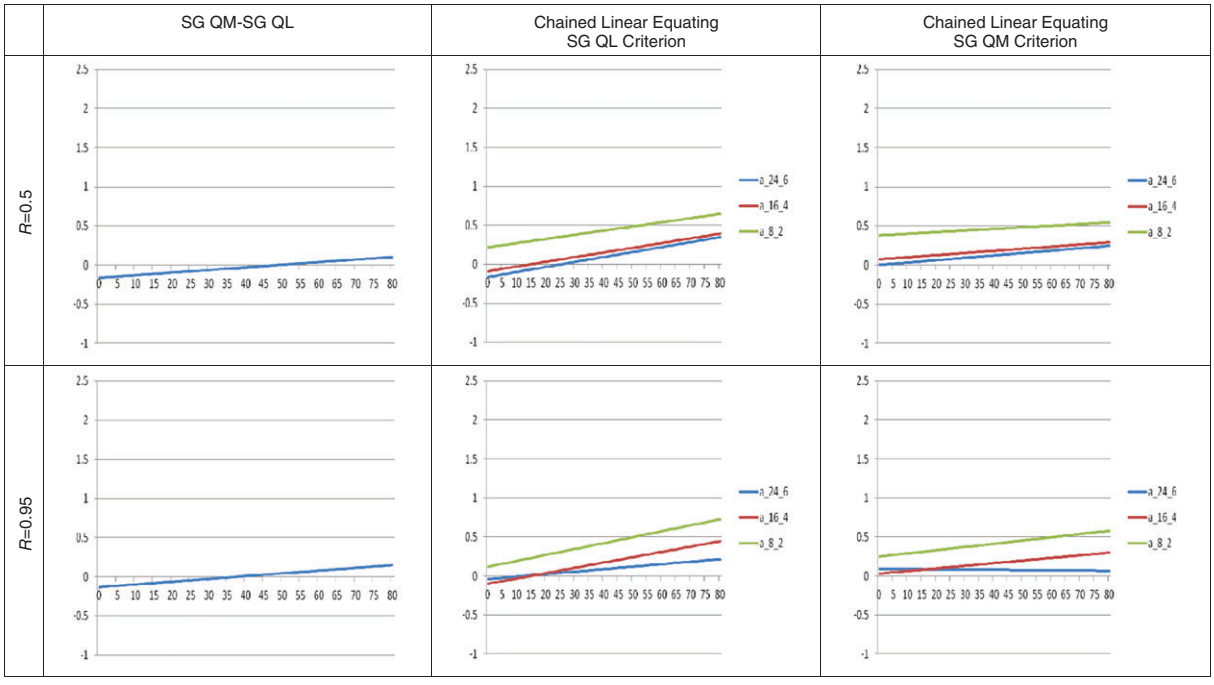

Figure A5 Same structure different difficulty $Y_e(4{:}1) \geq Y_m(4{:}1)$. Differential ability difference QL QM. Chained linear equating.

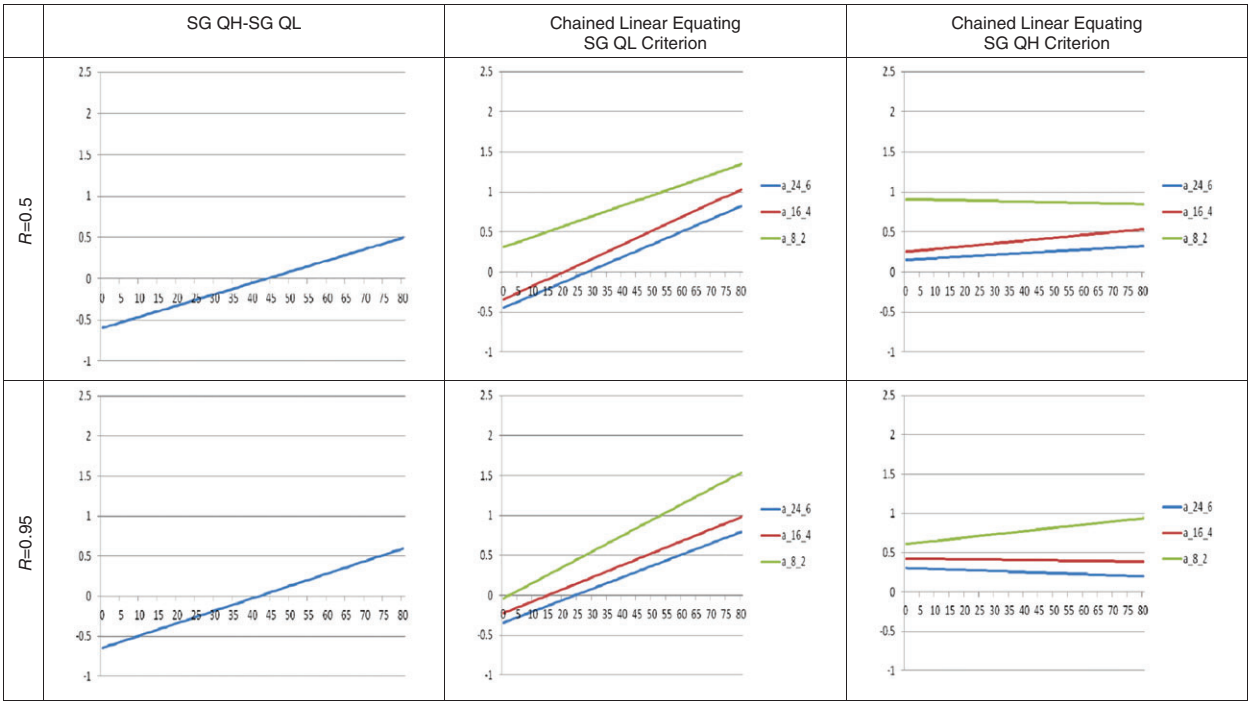

Figure A6 Same structure different difficulty $Y_e(4{:}1) \geq Y_h(4{:}1)$. Common ability difference QL QH. Chained linear equating.

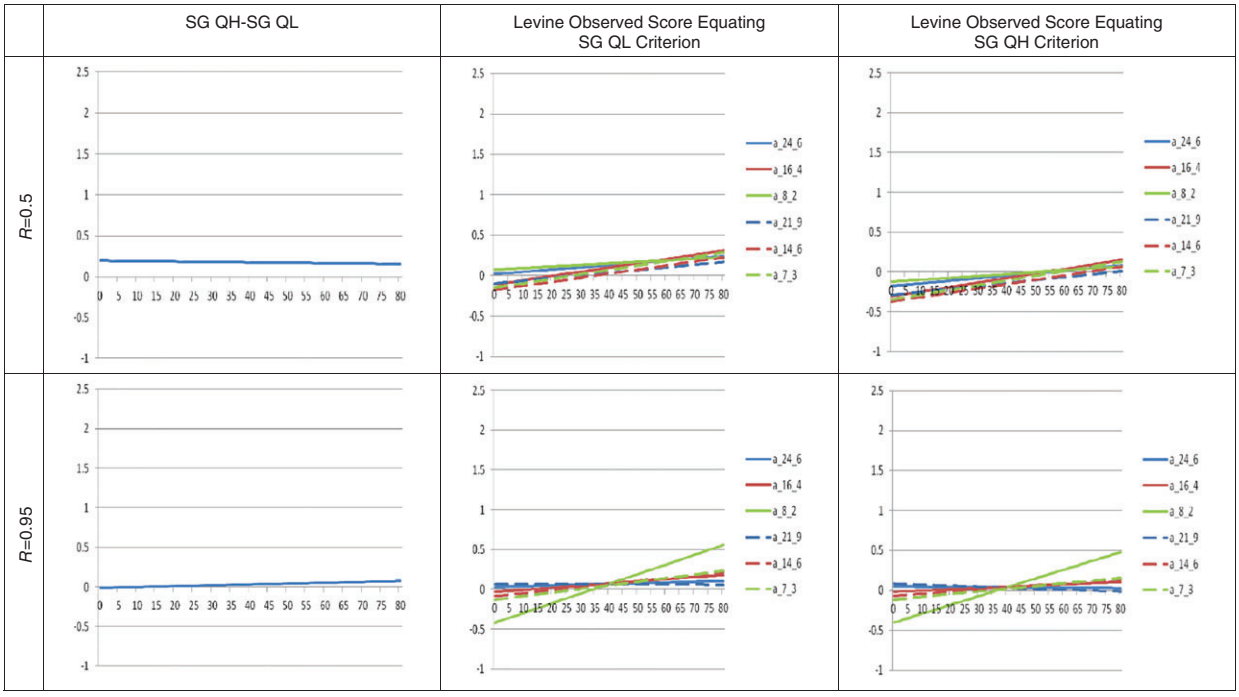

Figure A7 Different structure same difficulty $Y_e(4{:}1) \geq Y_e(3{:}2)$. Common ability difference QL QH. Levine observed-score equating.

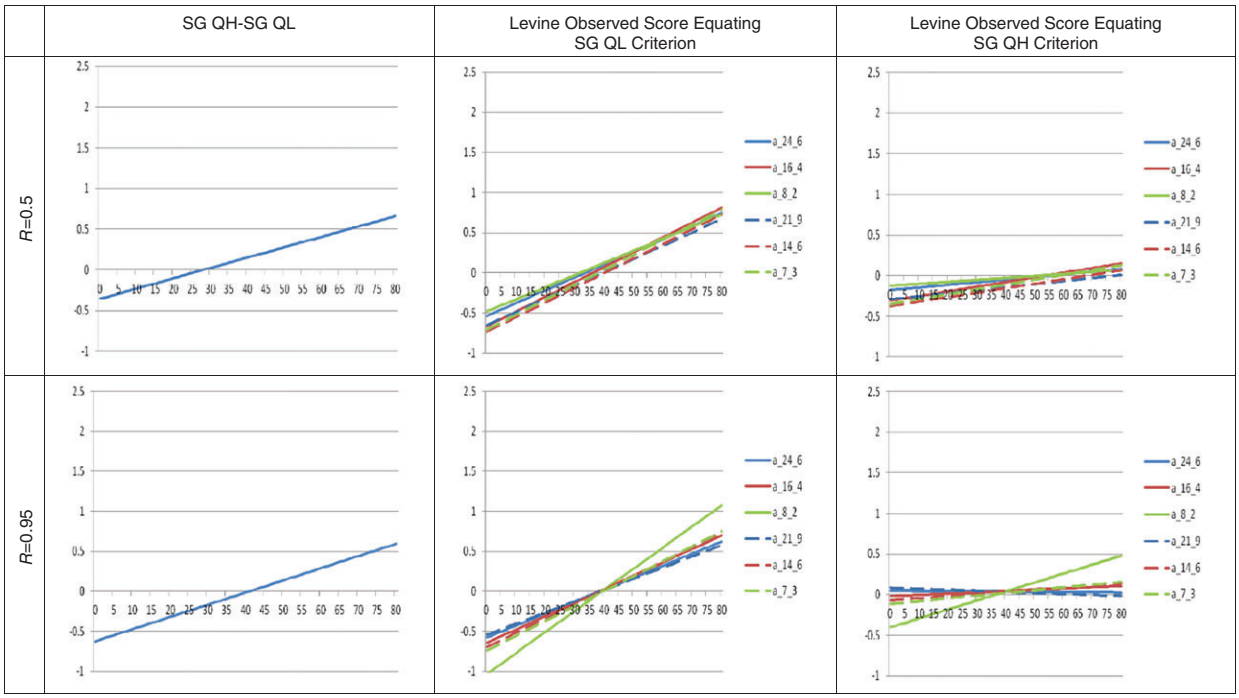

Figure A8 Different structure different difficulty $Y_e(4{:}1) \geq Y_h(3{:}2)$. Common ability difference QL QH. Levine observed-score equating.

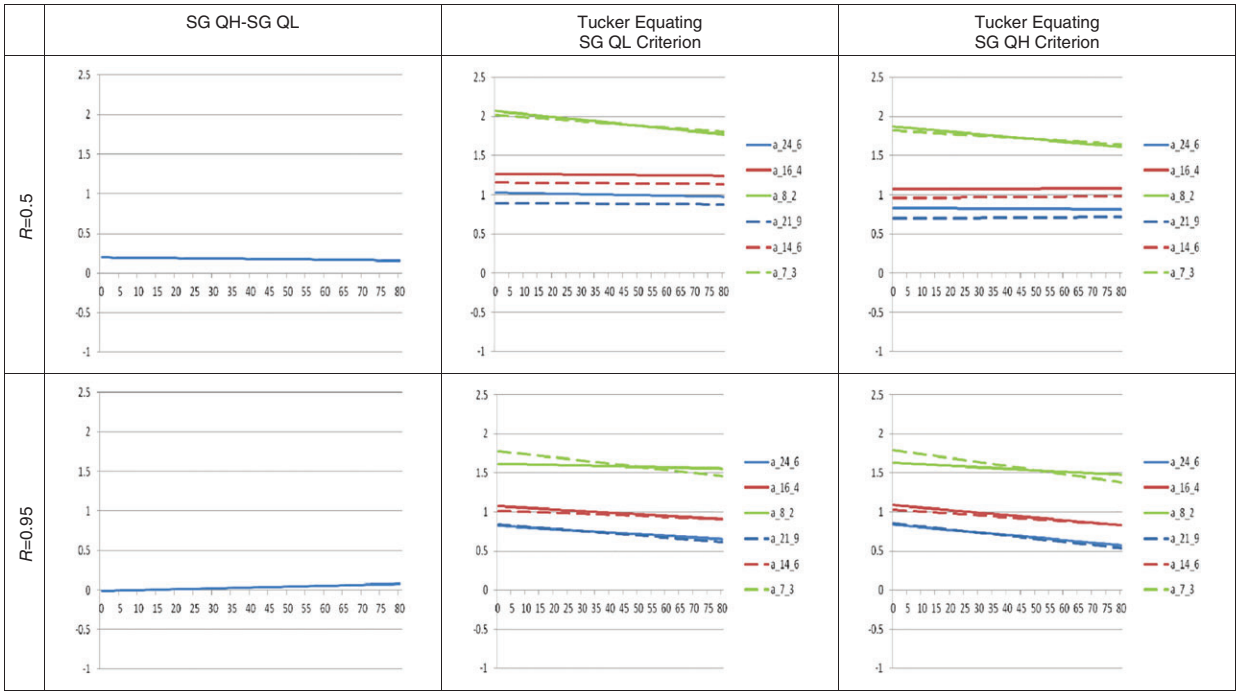

Figure A9 Different structure same difficulty $Y_e(4{:}1) \geq Y_e(3{:}2)$. Common ability difference QL QH. Tucker equating.

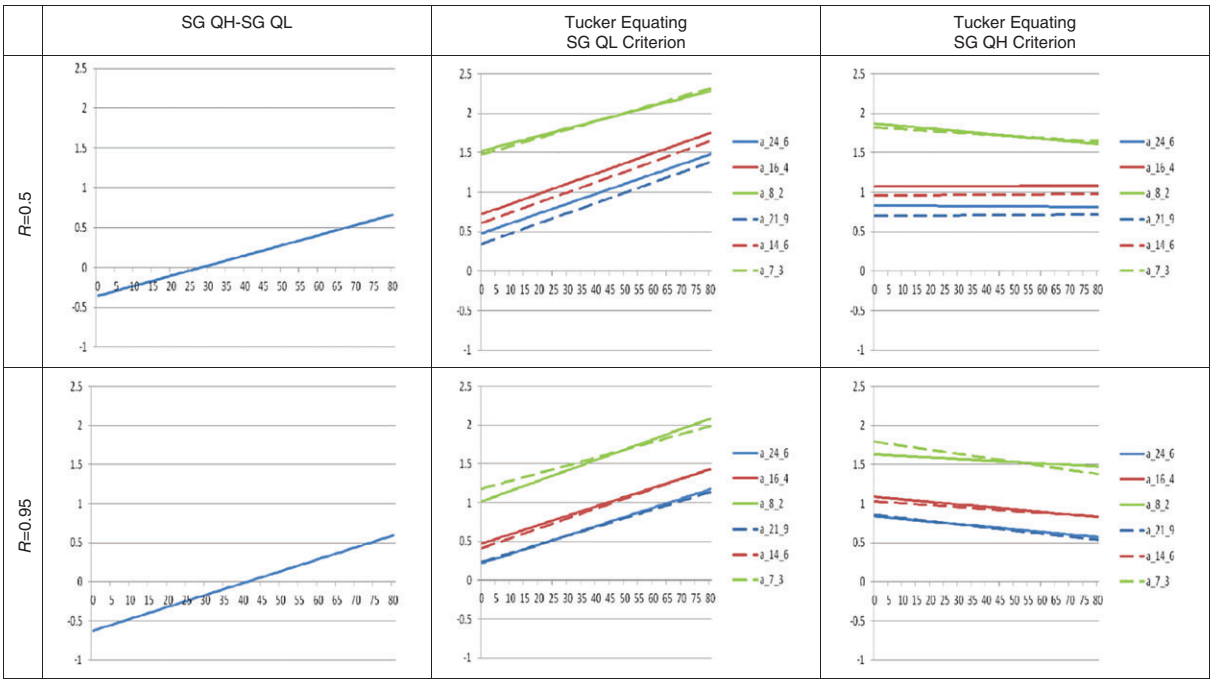

Figure A10 Different structure different difficulty $Y_e(4{:}1) \geq Y_h(3{:}2)$. Common ability difference QL QH. Tucker equating.

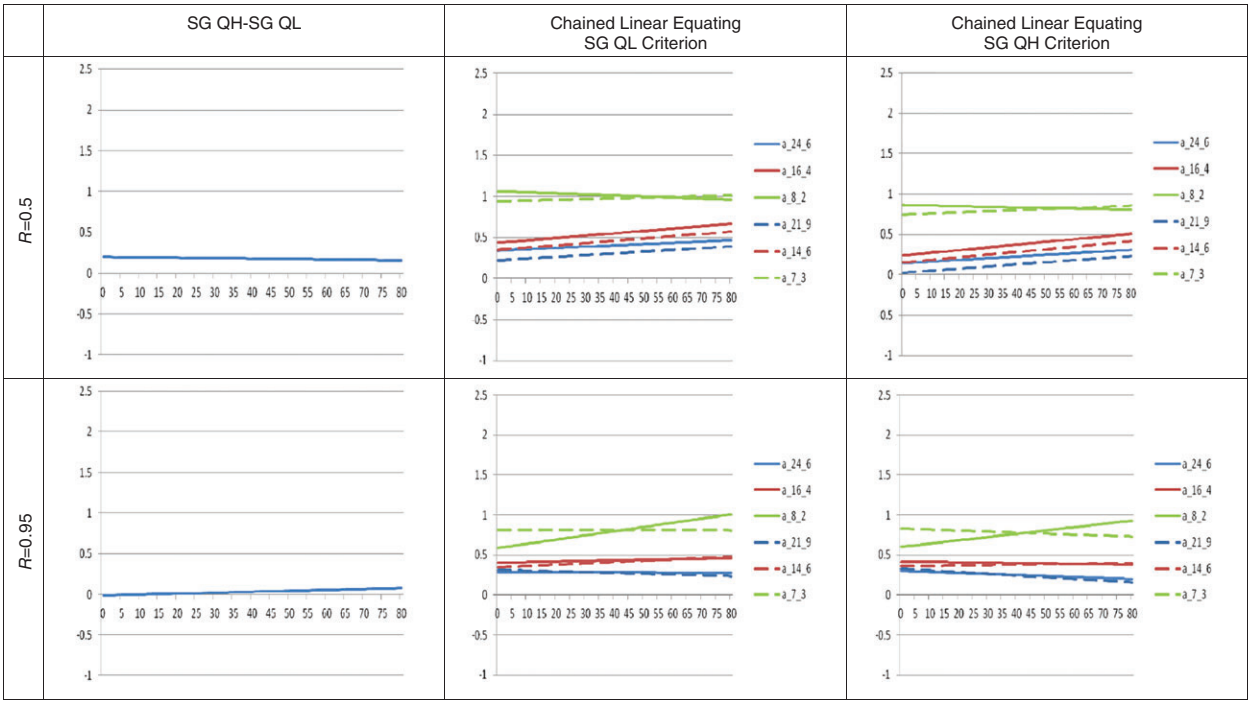

Figure A11 Different structure same difficulty $Y_e(4{:}1) \geq Y_e(3{:}2)$. Common ability difference QL QH. Chained linear equating.

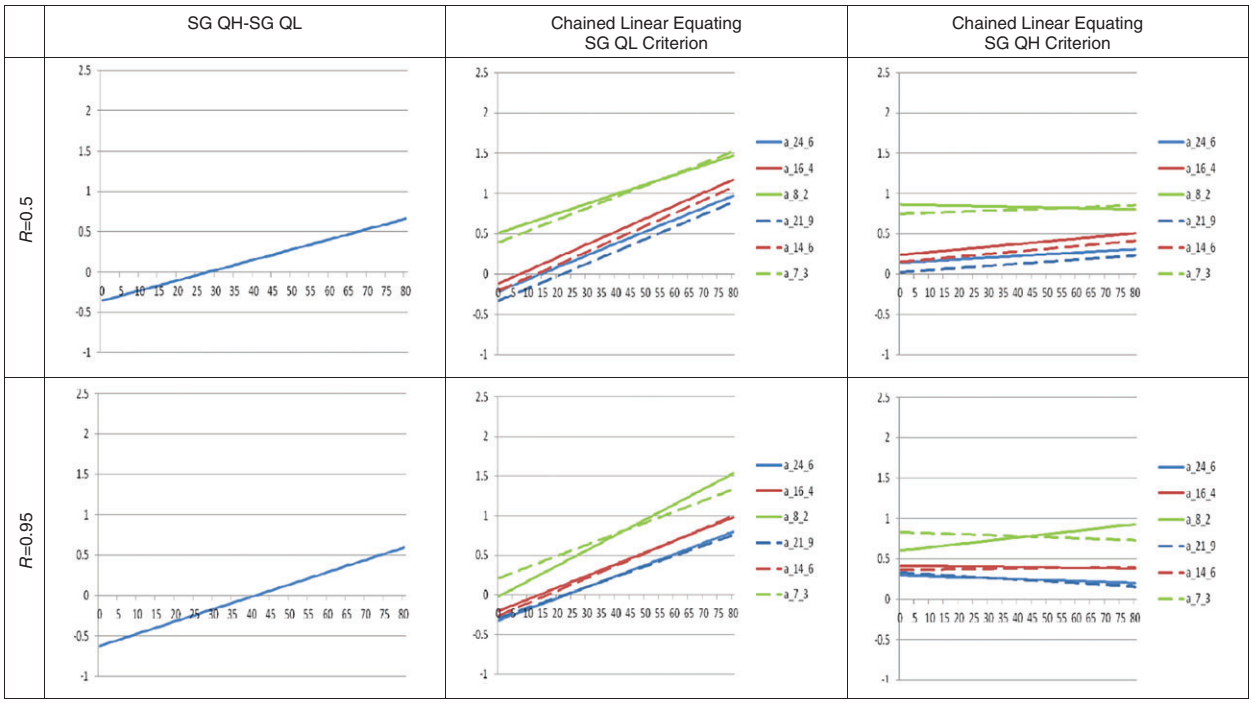

Figure A12 Different structure different difficulty $Y_e(4{:}1) \geq Y_h(3{:}2)$. Common ability difference QL QH. Chained linear equating.

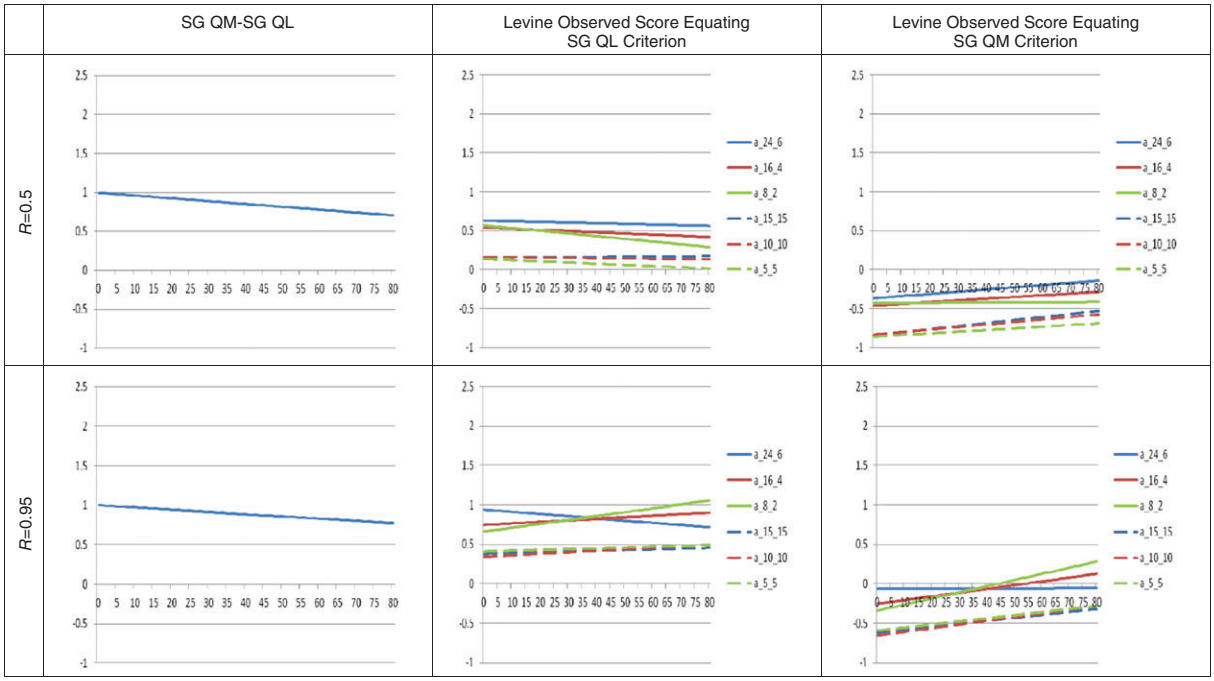

Figure A13 Different structure same difficulty $Y_e(4{:}1) \geq Y_e(1{:}4)$. Differential ability difference QL QM. Levine observed-score equating.

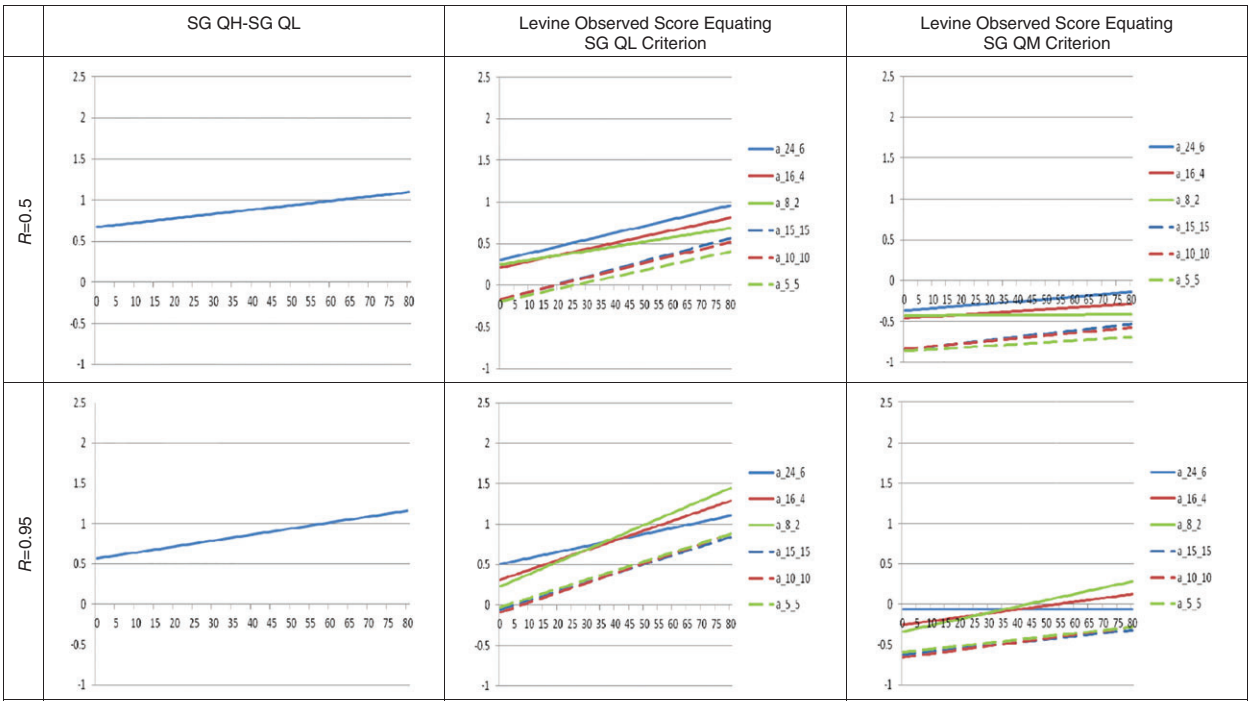

Figure A14 Different structure different difficulty $Y_e(4{:}1) \geq Y_m(1{:}4)$. Differential ability difference QL QM. Levine observed-score equating.

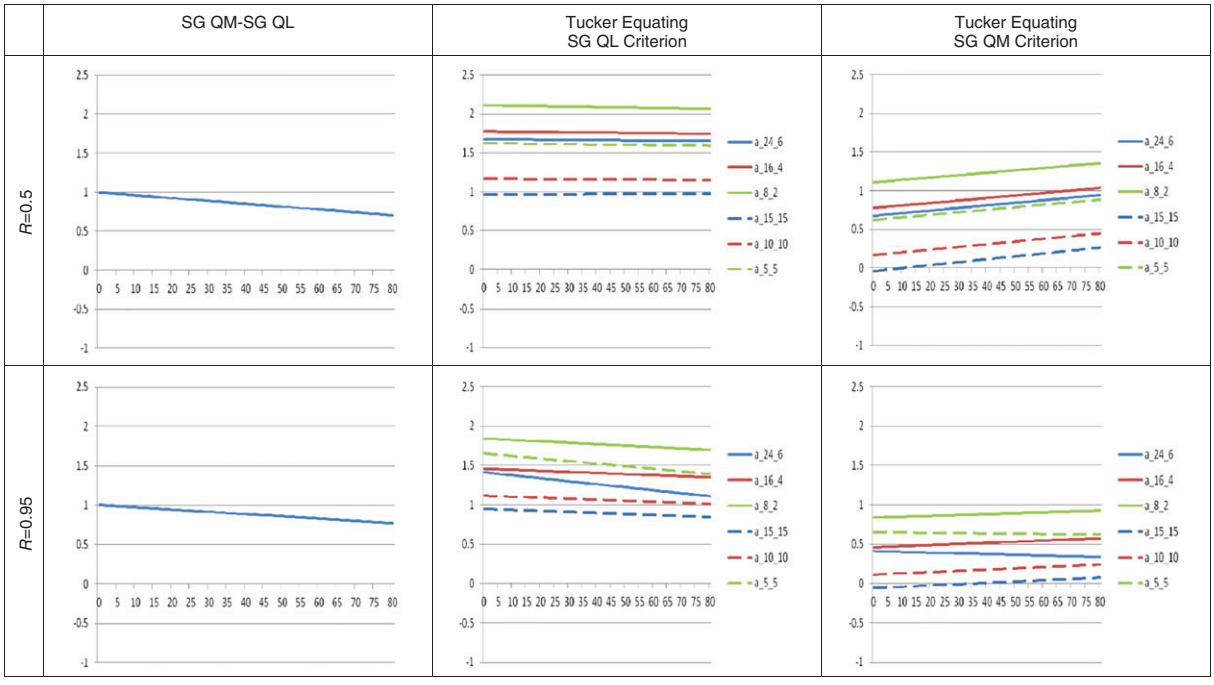

Figure A15 Different structure same difficulty $Y_e(4{:}1) \geq Y_e(1{:}4)$. Differential ability difference QL QM. Tucker equating.

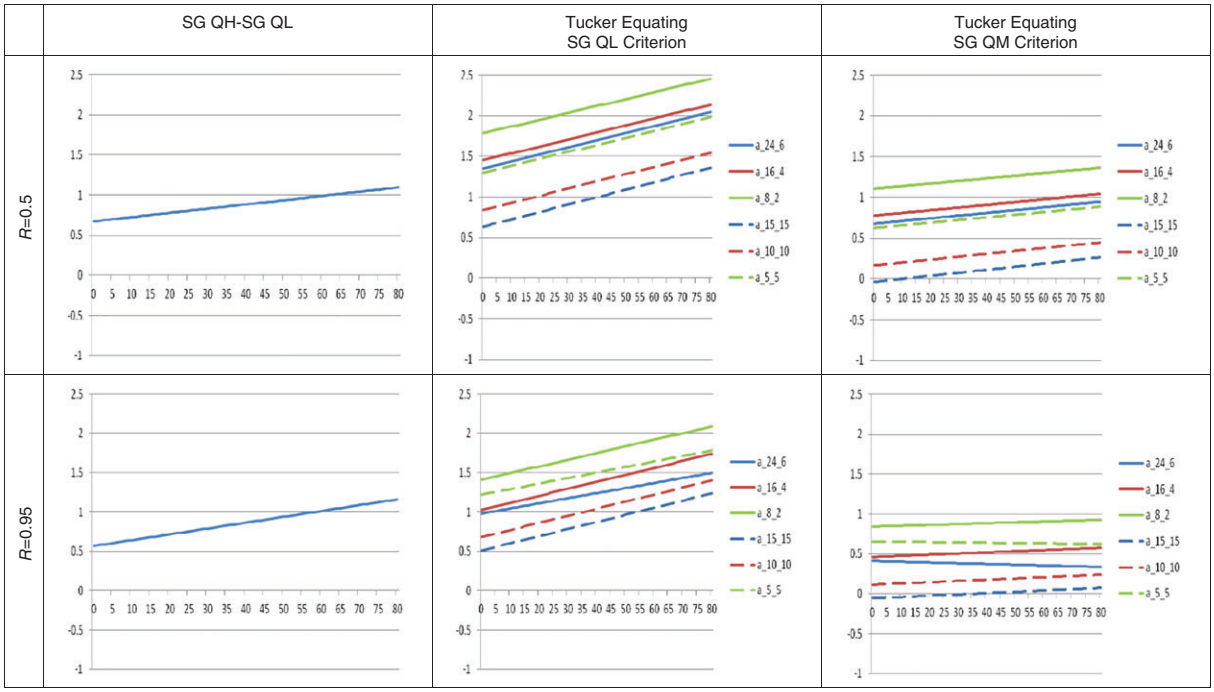

Figure A16 Different structure different difficulty $Y_e(4{:}1) \geq Y_m(1{:}4)$. Differential ability difference QL QM. Tucker equating.

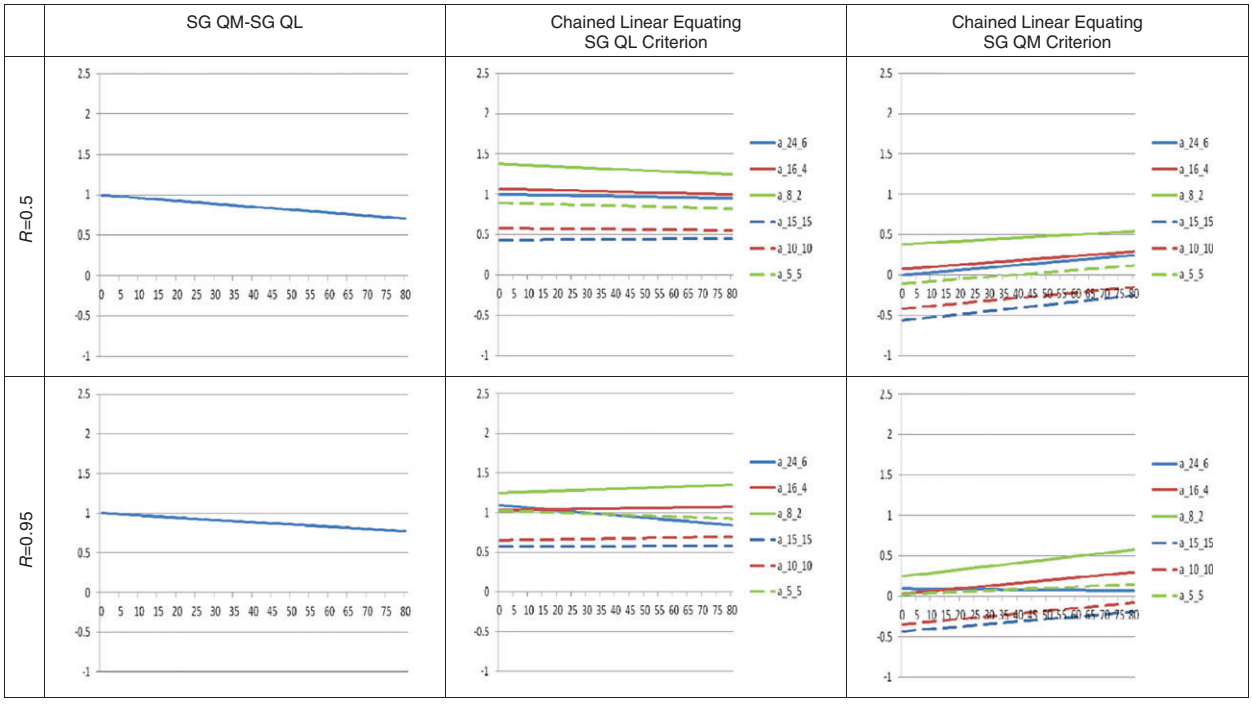

Figure A17 Different structure same difficulty $Y_e(4{:}1) \geq Y_e(1{:}4)$. Differential ability difference QL QM. Chained linear equating.

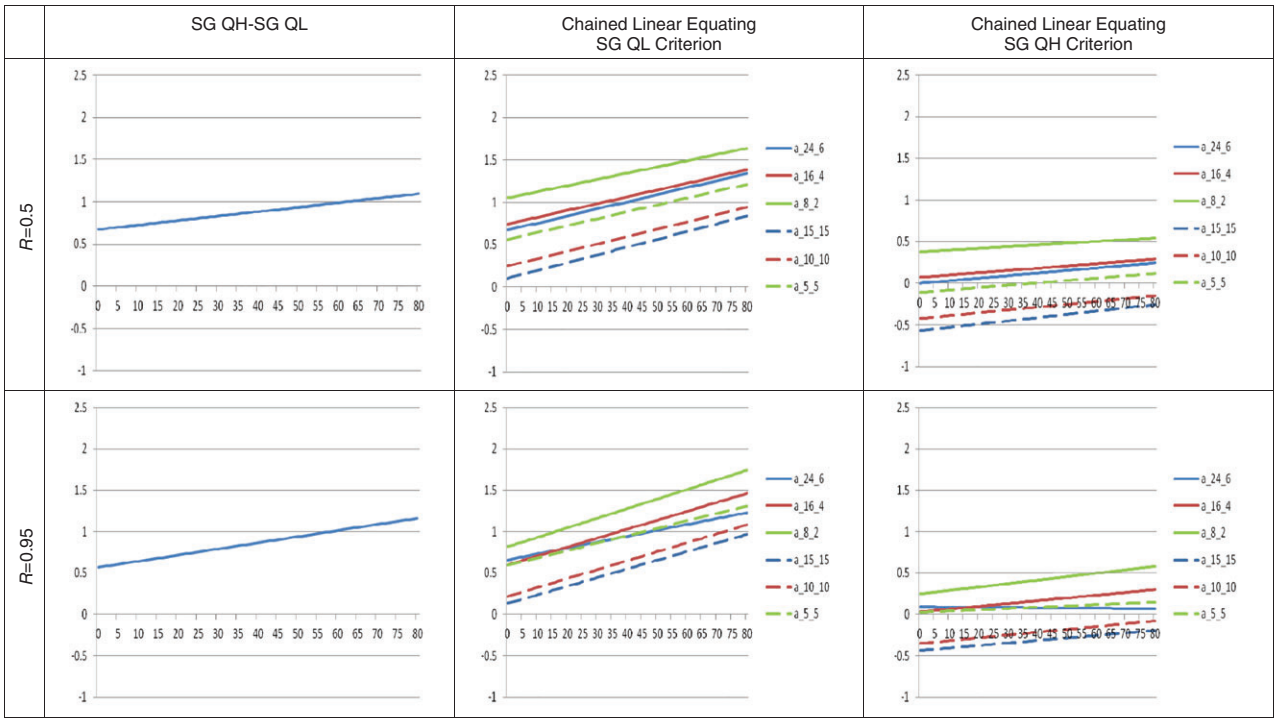

Figure A18 Different structure different difficulty $Y_e(4{:}1) \geq Y_m(1{:}4)$. Differential ability difference QL QM. Chained linear equating.

Suggested citation:

Lin, P., Dorans, N., & Weeks, J. (2016). *Linking composite scores: Effects of anchor test length and content representativeness* (Research Report No. RR-16-36). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12122

Action Editor: Gautam Puhan

Reviewers: Peter Van and Jiyun Zu