Listening. Learning. Leading.®

Research Report
ETS RR–11-04

Assessing the Falsifiability of Extreme Linking

Kyndra Middleton

Neil J. Dorans

February 2011

Assessing the Falsifiability of Extreme Linkings

Kyndra Middleton
Howard University, Washington DC[1]

Neil J. Dorans
ETS, Princeton, New Jersey

February 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Daniel Eignor

Technical Reviewer: Samuel A. Livingston

Abstract

Extreme linkings are performed in settings in which neither equivalent groups nor anchor material is available to link scores on two assessments. Examples of extreme linkages include links between scores on tests administered in different languages or between scores on tests administered across disability groups. The strength of interpretation attached to a linkage depends on the proper design and execution of a sound data collection plan. The current paper uses a real data set to illustrate how to indirectly assess the quality of linking the scores on two assessments that contain neither equivalent groups nor common anchor material.

Key words: extreme linking, falsifiability, population invariance, score equity assessment

Table of Contents

List of Tables

Page

1. Motivation

Substantial diversity among testing practices exists within the American educational system. Students are administered different standardized tests based on their state and even district of residency, and these tests are used to make critical decisions about students' academic careers. Recent demographic changes and a federal law that requires the reporting of test scores of all students have led researchers to examine the psychometric properties of tests that are administered to increasingly diverse subpopulations of examinees. For example, there has been a large increase of students who are not native English speakers in the United States population, and these students, referred to as *English-language learners* (ELLs), must nonetheless be included in score reports, even though the test scores of ELLs may not be interpreted in the same way as their native English-speaking counterparts' test scores for reasons that will be explained below.

When tests have different examinee subpopulations, or differ in construct or difficulty, test users are unable to make a straightforward interpretation of the test score. In these instances, one of several forms of linking can be performed to place the different tests on the same scale. If, however, subpopulations are different and there are no common items between the two tests, we face an extreme linking situation. The term, extreme linking, first introduced by Dorans, Pommerich, and Holland (2007), refers to linking scores on tests in situations in which neither equivalent groups of people nor truly common test material are available to separate group differences from test differences.

Without equivalent groups or common items between tests, there is no direct way of examining whether the linking is defensible. In essence, the link is not falsifiable. Falsifiability (Popper, 1959) is the logical possibility that an assertion can be shown to be false by an observation or a physical experiment. That something is *falsifiable* does not mean it is false; rather, it means that it is capable of being refuted. In this paper, we maintain that the nonfalsifiable problem of demonstrating that test forms across both measurement conditions and populations are equivalent can be indirectly, albeit inconclusively, assessed with a set of falsifiable questions that draw us closer to an evaluation of equivalence.

The current paper uses Popper's principle to illustrate how to handle extreme linkages. To begin, Section 2 delves into extreme linking in detail. In Section 3, we advance population invariance as a tool for assessing potential approaches to dealing with extreme linking data collection conditions. We then consider an extreme linking example in Section 4. We conclude in Section 5 with issues faced when conducting extreme linkages and a discussion of additional examples of where the falsifiability principle may be used.

2. Extreme Linking

As previously stated, *extreme linking* refers to linkings of test scores in situations in which neither equivalent groups nor anchor material is available. Linking the scores on tests given to students who use English as a second or foreign language (ELLs) to scores on tests given to students who use English as a first language is a prevalent example of an extreme linkage. These two subpopulations include two nonequivalent groups (examinees who are native English speakers and examinees who are ELLs) who are administered what appears to be the same content under different conditions of measurement (ELLs may receive an accommodation such as a read aloud, extended time, or even a bilingual dictionary; Native English speakers receive no accommodation). In this case, it would be impossible to have an anchor that will suffice for both groups. Native English speakers will not be able to complete the items with the same level of knowledge and understanding as the ELLs if the items are in a language other than English (i.e., the ELLs' language), and ELLs will not be able to complete the items with the same level of knowledge and understanding as the native English speakers if the items are in English with no accommodation provided.

Some words might not be readily translated from one language to the other in ways that would not alter the test's content. Assessing issues like differential reliability and differential test difficulty can be problematic when examinees who speak different languages are confronted by questions expressed in different languages and given under different conditions. It may be possible, however, to garner indirect evidence that sheds light on the type of linking that is feasible between scores on tests given in different

languages or administered under different testing conditions to groups who are conversant in different languages or assessed under different conditions.

Let X_{cg} represent a test given under condition *c* to group *g*, and let X_{dh} represent the same collection of items administered under condition *d* to group *h*. The premise of studies that employ *differential item functioning* (DIF) procedures to assess test adaptation, for example, is that scores on a matching variable X_{cg} mean the same thing as scores on a matching variable X_{dh}. This assumption of equivalent matching variables cannot be tested directly, however, since not only do the groups differ, but also the measurement conditions differ.

In this paper, we demonstrate how we can indirectly evaluate the assumption about the equivalence of X_{cg} to X_{dh} by looking at linking scores on X_{cg} to scores on X_{dg} and X_{ch} to X_{dh}. We can also examine the linking of scores on X_{cg} to scores on Y_{cg} (where *Y* represents test form *Y*) and compare it to the linkings obtained under condition *c* in group *h* and under condition *d* in both groups *g* and *h* to assess whether the forms are comparable.

Extreme linkages can be indirectly evaluated by employing *score equity assessment* (SEA), which was introduced by Dorans (2004). SEA can assess whether the linking relationship between scores on two tests is essentially invariant across populations and measurement conditions. The use of SEA in the current study demonstrates how SEA can be used in extreme linking situations. SEA utilizes correlation matrices and reliability coefficients to aid in addressing score level invariance.

3. Factors Affecting the Quality of Linking

There are several important conditions that affect the quality of linkings. These conditions have particular implications for extreme linkages.

3.1 Kolen's Features of Test Administration

Kolen (2007) discusses three features of test administration that determine the type of linking that can be performed between scores on two different tests:

1. *Content similarity* of the two tests

2. *Administrative conditions* under which the tests are given

3. *Characteristics of the populations* administered the tests

The tests' content should be similar, built to the same specifications; the administration conditions should be the same; and the characteristics of the populations administered the test should be similar in order to achieve linking.

Content (1.) is very important for assessing extreme linkages because a test's content is related to the test's construct and thus plays a large role in an examinee's test score. If the content is changed significantly across test forms or versions of a test, the linking function will be affected.

Administration conditions (2.) are often ignored or taken for granted yet they play an important role. The timing, the scoring instructions, the format of the test, and the mode of administration are all administration conditions. Is the language of the exam an administration condition as well? If we change the language in which a test is administered, have we changed the content of the test? Or have we changed administration conditions, or both the content and administration conditions?

The characteristics of the populations (3.) also are very important when it comes to assessing the extreme linkage of scores on tests administered to distinct populations, e.g., a Spanish-speaking sample and an English-speaking sample. If we administer a Spanish test to native speakers of Spanish and to an English-speaking group that also speaks Spanish, the content has remained the same but the change in population may have created an extreme linkage situation.

3.2 Invariance of Items, Anchors, and Tests

When faced with difficult challenges, there is a temptation to rely on strong assumptions in addressing the challenges. The strong assumptions of *item response theory* (IRT) have been employed by those facing extreme linkages (Dorans, 2007). If an IRT model fits the item data, item parameter invariance holds. *Item parameter invariance* occurs when the item parameters are not dependent on the sample of examinees used to calibrate the items. If item parameter invariance holds, the scores on a test administered in Spanish to Spanish-speaking examinees can be linked to scores on the same set of

questions administered in English to English-speaking examinees. If an IRT model fits the data, invariance would also hold for tests administered with and without accommodations. The assumption of item parameter invariance is very powerful, but it is sometimes violated even in normal linking settings. Therefore, this assumption is highly unlikely to be met in extreme linkage conditions. What we are talking about here is content invariance. Content invariance is the extent to which all the material on both tests remains similar. Scores on tests that exhibit content invariance when translated and back-translated are presumed to possess an identity linking relationship.

A weaker assumption than item parameter invariance is that the linking relationship between scores on Z_{spa} (Spanish version of the questions) to those on Z_{eng} (English version of the questions) is an identity. Note this identity function assumption does not require population invariance at the item level; it just assumes that the two tests are equivalent at the test score level. Individual items may or may not retain the same level of difficulty; the set, however, does. Another type of invariance is *condition invariance* which presumes, in effect, that the relationship between a set of items administered under one set of conditions (language or test administration) and the same set of items administered under another set of conditions (different language or different administration conditions) is an identity function. This is another strong assumption.

The use of common test material as an anchor is based on the same invariance assumption that underlies the use of the identity function for the entire test. It is reasonable to expect a test in one's own language to be easier and more reliable than a test in another language. This fact makes it virtually impossible to compare groups who take tests in different languages to each other via an anchor-test design that uses a direct linkage. As a result of these differences in conditions of measurement, it is difficult to justify that the identity assumption holds for the two versions of the items, whether this assumption is made about all items or a select anchor set.

Direct linkages occur when scores on X can be placed on the same scale as scores on Y directly, without performing any additional statistical analyses to link the scores on the two forms. Direct linkages between tests administered to different populations under different conditions are unlikely to be achieved because anchor material, like other

material on the test, is likely to be altered by the change in measurement conditions. The very strong assumption of invariance at the item level is not plausible nor is the weaker assumption of an identity relationship between the two versions of the test (or anchor) administered to disparate populations under different conditions.

In the next section we illustrate how several different methods, in conjunction, can be used to obtain indirect evidence as to how well scores on tests can be linked to each other across different administration conditions. By using the falsifiability principle, we show how an inaccessible problem becomes more accessible through performing a series of falsifiable steps. This approach requires the availability of bilingual groups, groups that take both tests, with each taken under its own measurement conditions.

4. A Real-Data Example Using Extreme Linkage Designs

In the context of extreme linkage, SEA asks: "Is the relationship between two test scores the same across *different measurement conditions*, as it is in a *reference condition*?" The question asks whether changes in *content*, *administration conditions*, or *population* matter. In contrast to presuming invariance of item, anchor, or test across testing conditions, SEA focuses on the invariance of the linking relationship between two scores, those on X and those on Y. This invariance of linking relationship between scores on X and scores on Y is not tantamount to stating X or Y is unaffected by the change in a measurement condition. Instead it says that they are changed in much the same way. For example, when two tests, X and Y, are administered without accommodations to equivalent groups of examinees that do not need accommodations, the data can be used to determine a direct linking relationship between X and Y. Below, however, we use SEA, along with other more familiar methods, to illustrate how to replace the intractable problem of linking scores on test forms across measurement conditions and populations, e.g., X_a to Y_s, where subscript a represents the accommodated version and subscript s represents the standard version of a test, with a falsifiable problem.

4.1 Study Design

The data used in this study were collected during a 2005 ETS experimental study (Laitusis, Cook, Cline, King, & Sabatini, 2008) in which fourth- and eighth-grade

students with reading-based learning disabilities (RLD) and without reading-based learning disabilities (NLD) were administered the Gates-MacGinitie Reading Tests (GMRT) fourth edition reading comprehension subtest under accommodated (audio) and nonaccommodated (standard) conditions (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). Although the Laitusis et al. (2008) study included eighth-grade students, for illustrative purposes, the current focus is on the 1,181 fourth-grade students, of which there were 527 RLD students and 654 NLD students.

4.2 Methodology

As a result of each student being administered the test under both accommodated and nonaccommodated conditions, two forms (*X* and *Y*) needed to be used. Schools were initially assigned to receive either the read-aloud or standard condition first, but in order to maintain a demographic balance among the two testing conditions, the order of the condition was then changed for certain schools. Within each school, randomly selected students received either *X* or *Y* first, leading to a counterbalanced design by form (*X* or *Y*) and order (Session 1 or Session 2). This was necessary because each student took the test under both conditions. The study design is reproduced in Table 1.

The empirical portion of this analysis addresses the following questions:

1. Is the difficulty of Forms *X* and *Y* affected by the change in accommodation within groups?
2. Are the reliabilities of Forms *X* and *Y* affected by the change in accommodation?
3. Are examinees in the RLD and NLD groups ordered in the same way by Forms *X* and *Y* when they are given under different conditions?
4. Is the linking between Forms *X* and *Y* invariant across group (RLD and NLD) and accommodation condition (*s* and *a*)?

Table 1

Design for NLD and RLD Groups

	Form	
	Session 1	Session 2
Group: RLD		
1a	X_s	Y_a
1b	Y_a	X_s
2a	X_a	Y_s
2b	Y_s	X_a
Group: NLD		
3a	X_s	Y_a
3b	Y_a	X_s
4a	X_a	Y_s
4b	Y_s	X_a

Note. Subscripts *s* and *a* represent standard administration and audio administration, respectively. NLD = without reading-based disabilities; RLD = with reading-based disabilities.

4.3 Results

Is the difficulty of Forms *X* and *Y* affected by the change in accommodation within groups? To determine whether the accommodation changed the test's difficulty, we examined the mean score and standard deviation of *X* administered under standard conditions compared to the mean score and standard deviation of *X* administered under audio conditions within the NLD group. The same was done for *Y* within the NLD group as well as for *X* and *Y* in the RLD group. Examining whether there is a change in difficulty within groups and forms allowed us to determine whether *X* and *Y* are equitable.

The means and standard deviations of Forms *X* and *Y* can be found in Table 2. The means for the NLD groups (Groups 1 and 2) on the accommodated versions of both *X* and *Y* are slightly higher (by between 1 to 2 points) than on the standard versions. Note also that the standard deviations are lower on the accommodated versions. This may in part be due to a ceiling effect because there were only 48 questions on each form. The means for the RLD groups (Groups 3 and 4) on the accommodated versions of both *X* and

Y are much higher (by between 5 to 8 points) than on the standard versions. The standard deviations are only slightly lower on the accommodated versions. This difference in difficulty within groups and within forms reveals that scores using the same forms across conditions were different; *X* and *Y* are hardly equitable across conditions in the RLD group. To place the scores on the same scale, some form of linking is needed.

Are the reliabilities of Forms *X* and *Y* affected by the change in accommodation? Table 2 also contains internal consistency reliabilities for both *X* and *Y* under both testing conditions (standard and accommodated). Examination of the reliability information in Table 2 reveals that both *X* and *Y* have reliabilities of .89 to .91 in the NLD group, and .88 to .89 in the RLD group. Based on these values, the effect of the accommodation on reliability seems to be small.

Table 2

Descriptive Statistics for NLD and RLD Groups—Same Examinees

Group	1	2	3	4
Number of examinees	328	326	269	258
Learning disability?	No	No	Yes	Yes
Form *X*	X_s	X_a	X_s	X_a
Condition	*s*	*a*	*s*	*a*
Mean	32.01	33.24	19.72	26.99
SD	9.34	8.27	8.91	8.80
Rel	0.91	0.89	0.88	0.88
Form *Y*	Y_a	Y_s	Y_a	Y_s
Condition	*a*	*s*	*a*	*s*
Mean	32.44	30.08	24.36	19.18
SD	8.81	9.68	8.81	9.05
Rel	0.90	0.91	0.88	0.89
Correlation of *X* and *Y*	0.78	0.78	0.51	0.61
Disattentuated correlation of *X* and *Y*	0.86	0.87	0.58	0.69

Note. Subscripts *s* and *a* represent standard administration and audio administration, respectively.

Are examinees in the NLD and RLD groups ordered in the same way by Forms *X* and *Y* when they are given under different conditions? To determine the strength of the relationship between *X* and *Y* under different conditions, the accommodated version and the nonaccommodated version were correlated. For the NLD group, examinees who received *X* under nonaccommodated conditions and *Y* under accommodated conditions were included in one correlation. The same was done for examinees who received *Y* under nonaccommodated conditions and *X* under accommodated conditions, resulting in two different correlations. This was then repeated for the RLD group, resulting in an overall total of four correlations. Additionally, the reliability coefficients were used to correct these correlations for attenuation due to measurement error.

Correlations, both observed score and true-score, are also presented in Table 2. The bottom of the table reveals that the correlation between scores obtained under standard conditions with those obtained under audio conditions is .78 in the NLD groups, which corrected for attenuation is about .87. In the RLD groups, the correlation of *X* and *Y* scores for Group 3, which received *X* standard and *Y* accommodated, is .51 (.58 disattenuated). In Group 4, the correlation of *X* to *Y*, with *X* being the audio version and *Y* being the standard version, is .61 (.69 disattenuated). These differences in disattenuated correlations within the RLD group demonstrate that the accommodation clearly has an effect on the construct measured in this group as can be seen by the lower correlations. This different also shows that the RLD group is not ordered in the same way by forms *X* and *Y* when the forms are given under different conditions.

Is the linking between Forms *X* and *Y* invariant across group (RLD and NLD) and accommodation condition (*s* and *a*)? Because of the small sample sizes, we employed a linear equating (mean sigma) method to convert scores from the raw score scale of test form *X* to the scale of form *Y*. We presumed equivalence among the two RLD groups as well as among the two NLD groups and used the mean sigma method in each of the four groups. Linking invariance was tested by examining both the *mean difference* (MeanDiff) and the *root expected square difference* (RESD). The MeanDiff is

simply a weighted average of differences between a subgroup linking function and the reference group linking function:

$$MeanDiff = \frac{\sum_{i=0}^{I} N_{ig}\left[\left(a_g - a_r\right)X_i + \left(b_g - b_r\right)\right]}{N_g}$$

where *i* represents each raw score point; N_{ig} represents the total number of members in the subgroup *g* at the i^{th} score point; $\left(a_g - a_r\right)X_i + \left(b_g - b_r\right)$ is a difference in the linear linking functions that link scores on *X* to scores on *Y* with a_g representing the slope of the linear linking function for the subgroup *g*, a_r representing the slope of the linear linking function for the reference group *r*, X_i representing the i^{th} raw score point on *X*, b_g representing the intercept of the linear linking function for the subgroup *g*, b_r representing the intercept of the linear linking function for the reference group *r*; and N_g is the total number of examinees in the g^{th} subgroup.

The RESD is also an average that uses the same subgroup frequencies to weight and sum the squared differences at each score level. The last step is to take the square root of the average squared difference. To find the RESD, the following equation can be used:

$$RESD = \sqrt{\sum_{i=0}^{I} N_{ig}\left[\left(a_g - a_r\right)X_i + \left(b_g - b_r\right)\right]^2 / N_g}$$

Small values of MeanDiff and RESD, such as one that is less than half of a raw score point, indicate that the linking function between the subgroup and the reference group used as a comparison group, i.e., the NLD nonaccommodated group (Group 1 in Table 4), differs by less than a difference in unrounded raw scores that should convert to the same rounded raw score.

To answer the question above, *Is the linking between Forms X and Y invariant across group (RLD and NLD) and accommodation condition (s and a),* the data were

grouped into four subgroups based on disability status and accommodation, ignoring the order in which the test form was administered: (a) NLD nonaccommodated (NLD-s), (b) NLD accommodated (NLD-a), (c) RLD nonaccommodated (RLD-s), and (d) RLD accommodated (RLD-a). The order was ignored because when analyses were performed using Session 1 examinees only, the results using the MeanDiff and RESD indices were very similar to those obtained when Session 1 and Session 2 examinees were combined.[2] (Refer to Table A in the Appendix to see the data broken into groups based on forms and order. Order does seem to have an effect on the correlations as can be seen in Table A of the Appendix.)

Table 3 contains sample sizes, means, standard deviations and internal consistency reliabilities for both X^t and Y under both testing conditions, where X^t is the transformation of X onto the scale of Y via $a_r X_i + b_r$ obtained in the reference group NLD-**s** when both tests were administered under standard conditions. The table contains two additional statistics that summarize the similarity of equating results based on different subpopulations, MeanDiff and RESD. The mean scores for X^t in Table 3 are different and lower than the X mean scores in Table 2 because the X^t means in Table 3 have been converted to the Y scale on the basis of the reference group linking. Y is harder than X, as seen in the fact that each mean (Table 3) for X^t is about 2 points lower than the corresponding mean (Table 2) for X in each combination of condition (*s* or *a*) and group (NLD or RLD).

The second column at the bottom of Table 3 contains 0.00 for both MeanDiff and RESD because the reference group is defined to be the NLD group under standard conditions. This group was selected as the group for the reference condition linking because it is the group on which the linking of scores on X to scores on Y would be most likely to occur. The MeanDiff and RESD values summarize what would happen if the reference group linking was replaced by subgroup-specific linking. When the subgroup is the same as the reference group, as it is in the case of NLD-standard (NLD-s**)**, the two subgroup and reference conversions are the same, and the difference statistics are zero.

Note the direction of the difference, $\left(a_g - a_r\right) X_i + \left(b_g - b_r\right)$, is subgroup condition minus reference condition; a positive difference indicates that the reference group

conversion is lower than the subgroup conversion. This difference is easiest to understand if the slopes are the same in the reference condition and the other conditions because the difference in conversions is then the difference in additive constants between each nonreference condition linking and the reference condition linking. A positive difference means that compared to the *X/Y* difficulty relationship observed in the reference group, *Y* appears relatively more difficult compared to *X* in the subgroup.

Table 3

Descriptive and Invariance Statistics for NLD and RLD Groups That Took X and Y Under the Same Conditions on the Scale of X Based on the Linking in the NLD Group Under the Standard Condition, the Reference Condition Linking

Group	1		2		3		4	
Learning disability?	No		No		Yes		Yes	
Form	Y_s	X_s^t	Y_a	X_a^t	Y_s	X_s^t	Y_a	X_a^t
Number of examinees	326	328	328	326	258	269	269	258
Mean	30.08	30.08	32.44	31.36	19.18	17.34	24.36	24.88
SD	9.68	9.68	8.81	8.57	9.05	9.23	8.81	9.12
Reliability	0.91	0.91	0.90	0.90	0.89	0.89	0.88	0.88
MeanDiff	0.00		1.08		1.83		-0.52	
RESD	0.00		1.10		1.84		0.60	

Note. Subscripts *s* and *a* represent standard administration and audio administration, respectively. The superscript *t* indicates that *X* has been transformed to the *Y* scale based on the linking between *X* and *Y* in Group 1, the NLD group that received both tests under the standard condition. This is the reference condition linking.

The third column at the bottom of Table 3 suggests that use of the NLD-**s** reference conversion (31.36) in place of the NLD-audio (NLD-a) conversion in the NLD-a condition would result in means for *X* on the scale of *Y* that would be about 1.08 points lower than if the subgroup conversion (32.44) were used. In other words, under the accommodated condition, *X* appears easier relative to *Y* than it appears to be without the accommodation,

the reference condition. Invariance seems to be violated somewhat when the accommodation condition changes from accommodated to standard for the NLD group**.**

The results for the RLD groups in columns 4 and 5 of Table 3 also confirm that *Y* is harder than *X* for this group of students. It also shows that the RLD groups score lower than the NLD groups (compare columns 4 and 5 to columns 2 and 3). The effect of the accommodation for the RLD group is quite evident in row 5 of columns 4 and 5, where the accommodated means (column 5) are about five to seven points higher than the nonaccommodated means (column 4).

Column 4 at the bottom of Table 3 contains large values of both MeanDiff and RESD, about 1.83 points. These values show that use of the reference conversion, based on the NLD-s condition, in place of the RLD-standard (RLD-s) conversion in the RLD-s condition would result in a mean (17.34) for *X* on the scale of *Y* that would be about 1.83 points lower than the mean (19.18) obtained if the subgroup conversion was used. In other words, the linking of *X* to *Y* in the RLD-s condition attenuates the relative easiness of *X* compared to *Y* that is seen under the NLD-s reference condition. Invariance of the relationship between *X* and *Y* is clearly violated when the NLD group is replaced by the RLD group and the tests are administered under standard conditions.

Column 5 at the bottom of Table 3 contains the findings for the interesting case in which the conversion for the reference linking NLD-s is compared with the RLD-audio (RLD-a) conversion with the RLD-a group. Here both the population has changed and the administration conditions have changed, yet the results exhibit the strongest degree of linking invariance of any condition compared to the reference condition, as evident in the MeanDiff of -.52 and the RESD of .60. These values show that use of the reference, NLD-s, conversion in place of the RLD-a conversion in the RLD-a condition would result in a mean (24.88) for *X* on the scale of *Y* that would be about .52 points higher than if the subgroup conversion (24.36) were used. In other words, the linking of scores on *X* to scores on *Y* in the RLD-a condition makes the relative easiness of *X* appear only slightly more than it is under the traditional NLD-s condition. These results are relatively invariant.

Whenever there is a difference between MeanDiff and RESD, as in the RLD-a condition, it means the reference conversion and the subgroup conversion crossover

where there are enough data for the crossover to be noticed. This can be seen in Figure 1, which reveals that all difference lines are not parallel to the reference group line, NLD-s, and show where differences are large (at the higher score levels for all conversions and where they are small (at the lower score levels for RLD-a and NLD-a) . They are never small for the RLD-s case.

Invariance of the linking relationship between scores on two test forms did not hold when the forms were administered to the same group under different measurement conditions or when the tests were administered under standard conditions to the RLD and NLD groups. The illustration indicated, however, that the linking relationship between scores on two forms was closest to the reference condition when the test was accommodated to the group's disability. Note that this degree of linking invariance does not mean that X_a in the RLD group and X_s in the NLD group are equivalent measures. Rather it means that their difficulties relative to the difficulties of their corresponding Y scores were closer than what was observed between X and Y under the NLD-s condition and between X and Y observed under either the NLD-a condition or the RLD-s condition.

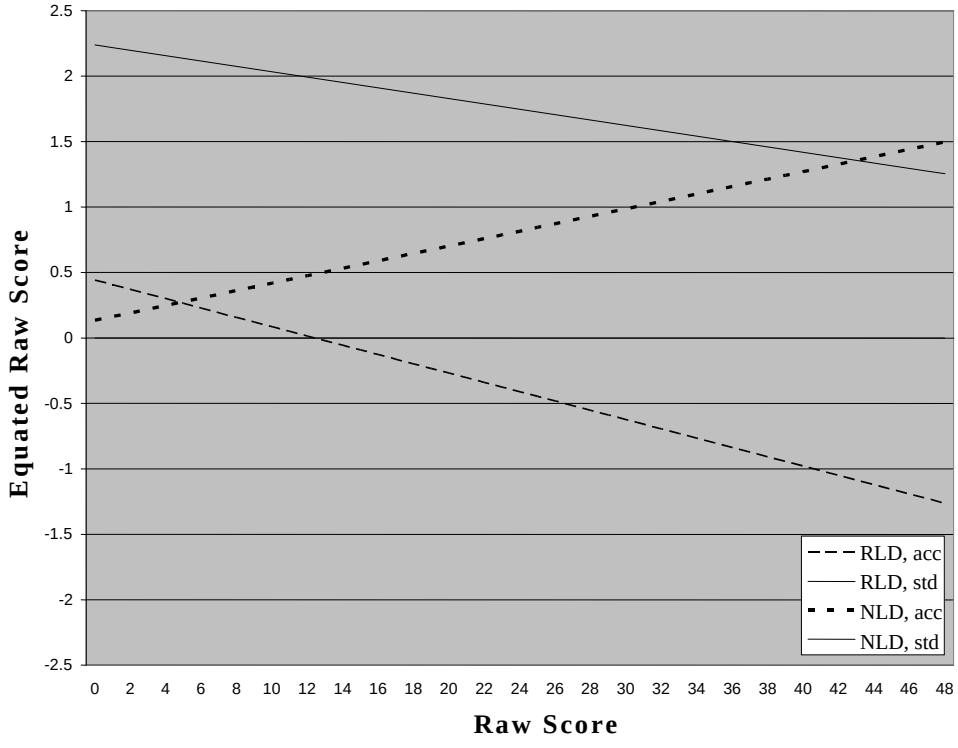

Figure 1. **Difference lines (subgroup conversion – NLD-s conversion).**

5. Discussion

Kolen (2007) discusses three features of test administration that determine the type of linking that can be performed between scores on two different tests: content similarity of the two tests; administrative conditions under which the tests are given, characteristics of the populations administered the tests. The test should have similar content, be built to the same specifications, be administered under the same conditions, and the characteristics of the populations administered the test should be similar in order to achieve solid linking. Extreme linkage may involve different content (items translated into another language), different measurement conditions (accommodated vs. standard), and almost always different populations (students with disabilities, [LD] vs. students without disabilities [non-LD]).

When faced with challenging situations like this, there is a temptation to lean on a strong model and let it provide an answer. Typically, the model's assumptions are violated because it presumes that item or test properties are unaffected by changes in population and conditions. Different types of invariance can be checked. These include (a) item level invariance such as item parameter invariance and (b) score level invariance such as invariance of reliability, and (c) invariance of linking function.

Extreme linking is often employed in several instances that involve nonequivalent groups and anchor items that appear to be common but in fact are not common among tests. One instance of extreme linking is in studies that compare how students in the United States fare on cognitive tests relative to other students in other nations. Although these comparisons involve students who are tested in their dominant national language, some of the testing material used for these comparisons are common in the sense that the questions have been translated from one language to the other and then back-translated to their original form. The presumption is that these common questions present the same task to the two different language populations in the sense that the questions are equally difficult in a special sense.

Evidence from the DIF literature challenges this presumption. Schmitt and her colleagues (e.g., Schmitt, Holland, & Dorans, 1993) demonstrated that items that had true cognates in Spanish that were used more frequently in Spanish than the words were used

in English were easier for Hispanics than for matched Whites. Would these items, which exhibit DIF in the English language, translate and back translate without a hitch? Would we expect DIF if they were administered in Spanish and we compared examinees who spoke Spanish as their first language to examinees who did not speak Spanish as their first language? If so, linking scores obtained from tests administered in different languages (*test adaptation*) or using different accommodations is a difficult challenge.

Research conducted by Cascallar and Dorans (2005) on linking the English language SAT® to the Spanish Language Prueba de Aptitud Acadèmica (PAA) demonstrates how tricky it is to link scores on tests by employing a bilingual group, perhaps because it is hard to conceptualize the process of thinking independently of language. Since the PAA is an exam used either for entry into Puerto Rican, Mexican, and other Latin American colleges and universities or as part of an admissions portfolio for students living in those countries who wish to attend a United States college or university, it is important that the PAA and the SAT be linked to ensure proper admission decisions are made. Cascallar and Dorans examined relationships between scores on the SAT and scores on the PAA in a bilingual group. The relationship that the English as a Second Language Achievement Test (ESLAT) had with the two verbal scores, SAT I-V and PAA-V, which were built to essentially the same content specifications, was revealing. The SAT-Verbal (SAT-V) and PAA-Verbal (PAA-V) correlated .62. The ESLAT had noticeably higher correlations with the SAT-V (.74) than did PAA-V. The ESLAT score was not that highly related with PAA-V score (.45). Clearly, the assessment of developed verbal ability is intimately tied to the language in which the verbal ability is assessed. These results suggest that under conditions of test adaptation score linking is not likely to yield equated scores.

In a study done by Allalouf, Hambleton, and Sireci (1999), 34% of the common verbal items on the Psychometric Entrance Test (PET) displayed DIF. Examinees received either the Hebrew version or the Russian version of the test with 40–44 items in common. Although the translation of the common items from Hebrew to Russian was intended to retain the same meaning and item difficulty level, this was not the case as evidenced by the amount of DIF items. In another study involving test translation, Gierl,

Rogers, and Klinger (1999) found that on a Social Studies Achievement Test that was translated from English to French, SIBTEST, Mantel-Haenszel, and logistic regression identified the same 19 items (out of 49) as having DIF. Substantive hypotheses could be generated for only six of those 19 items.

When dealing with accommodations, Bielinski, Thurlow, Ysseldyke, Freidebach, and Freidebach (2001) found that both reading comprehension and math items were more difficult for LD students who received a read-aloud accommodation than for LD students who did not receive a read-aloud accommodation, when compared to their non-LD peers. In reading, 24% of the 41 items in the nonaccommodated group displayed DIF compared to 46% of the items displaying DIF in the accommodated group. The math results were 3% and 9% of the 32 items for the nonaccommodated and accommodated groups, respectively. Just as with the Bielinski et al. (2001) results, a study done by Middleton (2007) found that a significantly larger amount of DIF was displayed in the accommodated group than in the nonaccommodated group. In both of these studies, LD students were given an accommodation where the accommodation was intended to decrease the difficulty for LD students and hence make the items more equal in difficulty for the two groups. In all these cases, the assumption was made that the common or anchor items were unaffected by the change in measurement conditions, whether it was language or accommodation. The preponderance of DIF suggested that the common items are not really common, that the matching variable used in the DIF analysis is suspect, and that the DIF analysis itself may not reflect DIF in the sense of DIF that is detected in a setting in which measurement conditions are held constant.

In the present study, which used the same data as the Middleton study, we have illustrated how to attack the intractable problem of linking scores on test forms across measurement conditions and populations, e.g., Y_a to $X_{s,}$ with a set of falsifiable problems. Statements about the equivalence of measures across different measurement conditions and populations without equivalent groups or anchor items are not testable. While equivalence across different measurement conditions and populations cannot be tested directly, it can be assessed indirectly by examining means, standard deviations, correlations, reliabilities, and linking relationships. If these analyses uncover evidence

that challenges the presumption that anchor items or matching material are common, as they did in our example, the results of the DIF analysis or score linking that had presumed some type of invariance, should be interpreted with proper caution. Testing the falsifiable in extreme linking settings should lead to the generation of fewer hypotheses about groups of examinees or items that do not hold with data collected under more controlled conditions.

The present research used a real data set from a carefully designed study to illustrate how assessing the falsifiability of invariance assumptions can be used to indirectly evaluate extreme linkings. Certain caveats are in order. First, simplicity of exposition, as well as small samples sizes, led us to use the linear equating model. Second, the assumption of random groups hinges on the degree to which the selected schools produced randomly equivalent groups. These two issues need to be addressed further before jumping to conclusions about these data.

References

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185–198.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cascallar, A. S., & Dorans, N. J. (2005). Linking scores from tests of similar content given in different languages: An illustration involving methodological alternatives. *International Journal of Testing, 5*, 337–356.

Dorans, N. J. (2004). Using population invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43–68.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research, 16, Supplement 1*, 85–94. doi 10.1007/s11136-006-9155-3

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). (Eds.). *Linking and aligning scores and scales.* New York, NY: Springer-Verlag.

Gierl, M., Rogers, W. T., & Klinger, D. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *Alberta Journal of Educational Research, 45*(4), 353–376.

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

Laitusis, C. C., Cook, L., Cline, F., King, T., & Sabatini, J. (2008). *Examining the impact of audio presentation on tests of reading comprehension* (ETS Research Report No. RR-08-23). Princeton, NJ: ETS.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-Macginitie reading tests, fourth edition*. Itasca, IL: Riverside Publishing.

Middleton, K. (2007). The effect of a read-aloud accommodation on items on a reading comprehension test for students with reading-based learning disabilities. *Dissertation Abstracts International*, *68*(9), 3818.

Popper, K. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating of hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Erlbaum Associates.

Notes

[1] This research was completed while the first author was a postdoctoral fellow at ETS.

[2] One of the drawbacks of the single-groups design listed above is the possibility of an order effect. Taking this into consideration, analyses were performed using Session 1 examinees only, and the results were very similar to those obtained when Session 1 and Session 2 examinees were combined. In the NLD group under same conditions, the MeanDiff for the accommodated group was 1.07 with an RESD of 1.09. In the LD group under same conditions, the MeanDiff for the standard group was 1.84 with an RESD of 1.85. For the accommodated group, the MeanDiff was -0.47 with an RESD of .57. The reliabilities for each group was the same as the respective reliabilities when Session 1 and Session 2 examinees were combined. Based on these results, the sessions were combined to better address the SEA questions.

Appendix

Descriptive Statistics for NLD and RLD Groups - Same Examinees

Table A1

Descriptive Statistics for NLD and RLD Groups - Same Examinees

Group	1a	1b	2a	2b	3a	3b	4a	4b
Number of examinees	159	169	166	160	137	132	122	136
Learning disability?	No	No	No	No	Yes	Yes	Yes	Yes
Form *X*: condition	X_s	X_s	X_a	X_a	X_s	X_s	X_a	X_a
Sequence	1st	2nd	1st	2nd	1st	2nd	1st	2nd
Mean	32.55	31.51	31.84	34.69	20.67	18.73	26.12	18.86
SD	9.06	9.60	8.67	7.60	9.14	8.59	8.86	8.76
Reliability	0.91	0.90	0.89	0.91	0.89	0.88	0.88	0.88
Form *Y*: condition	Y_a	Y_a	Y_s	Y_s	Y_a	Y_a	Y_s	Y_s
Sequence	2nd	1st	2nd	1st	2nd	1st	2nd	1st
Mean	32.21	32.65	28.45	31.78	25.42	23.27	27.76	19.46
SD	8.69	8.93	9.71	9.38	8.70	8.82	8.71	9.33
Reliability	0.89	0.92	0.91	0.87	0.87	0.87	0.89	0.88
Correlation of *X* and *Y*	0.78	0.78	0.83	0.72	0.49	0.53	0.69	0.54
Disattenuated correlation of *X* and *Y*	0.87	0.86	0.92	0.81	0.56	0.61	0.78	0.61