JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Summer 2019, Vol. 56, No. 2, pp. 452–472

Use of Adjustment by Minimum Discriminant Information in Linking Constructed-Response Test Scores in the Absence of Common Items

Yi-Hsuan Lee
Educational Testing Service
Shelby J. Haberman
Consultant
Neil J. Dorans
Educational Testing Service

In many educational tests, both multiple-choice (MC) and constructed-response (CR) sections are used to measure different constructs. In many common cases, security concerns lead to the use of form-specific CR items that cannot be used for equating test scores, along with MC sections that can be linked to previous test forms via common items. In such cases, adjustment by minimum discriminant information may be used to link CR section scores and composite scores based on both MC and CR sections. This approach is an innovative extension that addresses the long-standing issue of linking CR test scores across test forms in the absence of common items in educational measurement. It is applied to a series of administrations from an international language assessment with MC sections for receptive skills and CR sections for productive skills. To assess the linking results, harmonic regression is applied to examine the effects of the proposed linking method on score stability, among several analyses for evaluation.

Many educational tests use both multiple-choice (MC) and constructed-response (CR) sections to measure different constructs. Often, the MC portion contains far more items than does the CR part of the test. While it is common practice to use some common MC items across test forms for the purpose of equating test scores, it is unwise for security reasons to use common CR items. Thus, the MC section scores may be linked to previous test forms via common items, but the CR section scores cannot be linked via common items. In other words, a commonly used equating design in practice—the nonequivalent groups with anchor test (NEAT) design—is typically not applicable to the CR sections. In addition, operational test forms are usually not spiralled and population invariance typically does not hold across administrations. Thus, it is also impossible to link the CR section scores via the equivalent-groups design. To adjust scores on CR tests, research existed that evaluated the use of an external measure containing MC items that was not constructed to be parallel to the CR tests as an external anchor in the NEAT design (e.g., DeMauro, 1992). However, DeMauro concluded that the external MC test and the CR test to be linked did not measure the same skills, so no further score linking was conducted for the CR test. Kolen and Brennan (2014, p. 346) noted that the practical constraints have made linking of CR test scores impossible in many practical circumstances. To address this

long-standing issue, they also urged more research to answer the questions of how and when external measures can be used to link scores for CR tests.

This study responds to Kolen and Brennan's call for research. It focuses on assessments where the MC sections and CR sections are designed to measure different constructs, and the MC section scores can be equated via common items but the CR section scores cannot. Instead of using the MC section scores as external anchors for linking CR section scores in a NEAT design (e.g., DeMauro, 1992), the linking method proposed in this article uses equated test scores for the MC sections as matching variables to establish pseudo-equivalent groups across test administrations and then link the CR section scores directly via the equivalent-groups design. As will be discussed in the Methods section, the conventional poststratification equating in the NEAT design may be conceptualized as a special case of the proposed linking method that relies on stronger assumptions about the anchor test scores that are difficult to meet for the type of assessments we studied. In contrast, the proposed methodology is more flexible: examinees for the form to be linked are weighted by minimum discriminant information (Csiszár, 1975; Haberman, 1984) so that the weighted examinee score distribution for the equated sections has selected moments equal to those for a reference distribution of examinees. The weighted distribution of the CR section scores (or composite scores) may then be linked directly to the corresponding score distributions for the reference distribution by use of conventional equating methods for equivalent groups (Kolen & Brennan, 2014). The target population of an assessment may be defined as the reference population. For assessments with a moving target, empirical data from the most recent years may be used to construct the reference population for weighting and linking.

Some precedents exist for the proposed methodology. For example, to mitigate some of the practical constraints with the use of CR tests, practitioners often use mixed-format tests that contain both MC and CR items to make some form of linking possible (Kolen & Brennan, 2014, ch. 8.8.2). The MC portion and the CR portion of a mixed-format test are intended to measure the same content area. In AP® examinations, which are mixed-format tests, conventional equating methods are first employed to link the MC sections for different administrations and then employed to link composite scores for different administrations to their corresponding MC sections (Dorans, 2003). There are alternative approaches to the methodology used by AP, and studies have been conducted to examine different unresolved issues in linking of mixed-format tests (see, e.g., Hanson, 1993; Kim, Walker, & McHale, 2010; Tate, 2000). These approaches share a common feature: they use common items in the mixed-format tests to link the composite scores. They cannot be used to link test scores that do not contain common items. The proposed methodology has advantages over the methodology used by AP, as well as the other approaches, because (a) unrounded, continuous test scores can be used for all stages of analysis, (b) scores from multiple test sections that measure content areas different from the CR sections to be linked can be used to weight samples, (c) weighted distributions can be obtained at once for both CR sections and composite scores, and (d) a direct raw-to-scale conversion can be produced for the scores to be linked. More importantly, it can be used to link test scores that do not contain MC sections as a subset.

The method of pseudo-equivalent groups (Haberman, 2015) that employs minimum discriminant information has been used to weight examinees in a test administration so that weighted marginal distributions of background variables match marginal distributions of background variables in a reference sample. In his data example, Haberman (2015) constructed 16 categorical variables from a background questionnaire to which the examinees were asked to respond. The background variables had modest missing data, provided rich information about the examinees of the assessment under study, and involved groups that did not differ much on the construct measured by the test. These conditions made matching on background variables effective for constructing pseudo-equivalent groups. However, such rich background information rarely exists for large-scale educational assessments. For tests with both MC and CR sections, the use of equated MC test scores from content areas that differ from the scores to be linked is likely to provide a superior match compared to the use of background information, as the equated MC test scores are available for everyone taking the test and more predictive of the CR test scores than is background information. To date, modification of Haberman's approach has not been used to link CR scores.

This article presents an empirical study of such a linking procedure using 1 year of data from a large-scale international language assessment. The linking procedure is an innovative extension of Haberman (2015) to address the long-standing issue of linking CR scores across test forms in the absence of common items in educational measurement. It is presented in the Methods section. To provide a motivation for this research, a data example is provided before the Methods section. A description of general properties of minimum discriminant information adjustment (MDIA) provides an indication of potential use in linking test scores. Sample weights resulting from MDIA are then used in a nonparametric equipercentile equating procedure based on weighted grade functions. An application of the linking procedure to the data example is discussed in detail, including the analysis procedure and linking results. Evaluation of the linking results is considered at the administration level, the individual score level, and the subpopulation level. In addition, an evaluation tool involving harmonic regression (Lee & Haberman, 2013) is adapted to the data example to examine the effects of the proposed linking method on score stability. The Discussion section addresses the implications of results for testing practice and discusses possible directions for future research.

Data Example

This example involves 1 year of data from $T = 42$ administrations of an international language assessment delivered to $n_+ = 380,533$ examinees worldwide who were nonnative speakers of a single target language. For each administration t, $1 \leq t \leq T = 42$, the assessment reported $k = 4$ test scores from two equated MC sections, listening and reading, and two unlinked CR sections, speaking and writing. See Table 1 for the assessment structure under consideration. The listening and reading sections measured receptive skills using all dichotomous items, except for a couple of polytomous items per test form. There were common items across test forms so that equating scores via common items was possible. The scaled scores of

Table 1
Assessment Structure in the Data Example

	Listening	Reading	Speaking	Writing
Equated MC section	√	√		
Unlinked CR section			√	√

Table 2
Pearson Correlations Between (Unlinked) Section Scaled Scores

	Listening	Speaking	Writing
Reading	.815	.627	.758
Listening		.711	.760
Speaking			.741

Table 3
Pearson Correlations Between (Unlinked) Mean Administration Section Scaled Scores

	Listening	Speaking	Writing
Reading	.913	.776	.777
Listening		.846	.770
Speaking			.787

listening and reading can be considered continuous after equating. On the other hand, the speaking and writing sections measured productive skills using solely CR items, and they were scored according to rubrics. Due to security concerns, neither the speaking nor the writing section shares common items across test administrations; hence, their scores are not linked.

The proposed linking procedure was applied to speaking and writing raw scores, based on the observed equated listening and reading continuous scaled scores. The high correlations in Table 2 between individual section scaled scores are promising for using listening and reading scores to link speaking and writing scores; even more promising are the even higher correlations of administration means of section scaled scores in Table 3 (see the Methods section for more discussion of the importance of high correlations). In this application, data from the 42 administrations are combined to form the reference population, and the same reference population is used to weight samples and link scores for individual administrations.

Methods

MDIA is a general procedure for weighting samples so that sample weights applied to a study population satisfy constraints corresponding to a target population. Within the field of statistics, it provides a generalization of poststratification (Cochran, 1977, pp. 134–135) and raking (Deming & Stephan, 1940). In linking applications, scores on $T \geq 2$ test forms from T different test administrations of an

assessment are to be linked. The examinees from administration t are a random sample from study population t. For positive integers t and u no greater than T, the k-dimensional random vector $\mathbf{X}_{tu}$ with elements X_{jtu}, $1 \leq j \leq k$, represents k hypothetical test scores for form t that would be obtained by a randomly selected member of study population u. For the data example considered in the article, the assessment has four sections, so the vector $\mathbf{X}_{tu}$ has $k = 4$ dimensions with elements X_{jtu}, $1 \leq j \leq 4$, representing four section scores. The test scores may be discrete or continuous. For $t \neq u$, the test scores are hypothetical, because $\mathbf{X}_{tu}$ is only observable if $t = u$. Suppose that the target population of the assessment is a synthetic population U (von Davier, Holland, & Thayer, 2004), which is used as the reference population for all T administrations under consideration. The k-dimensional random vector $\mathbf{X}_{tU}$ is defined for test scores for form t on the reference population U. In the linking procedure under study, the distributions of the score vector $\mathbf{X}_{tU}$, $1 \leq t \leq T$, are used to link scores for the T test forms by equating methods associated with randomly equivalent groups.

Consider form t, $1 \leq t \leq T$. For a positive integer d and a function g on the set of k-dimensional random variables, the d-dimensional random vector $\mathbf{S}_t = g(\mathbf{X}_{tt})$ on study population t has elements S_{bt}, $1 \leq b \leq d$. The $\mathbf{S}_t$ defines the d linear constraints to be satisfied in the MDIA procedure for study population t. Assume for simplicity that each vector $\mathbf{S}_t$ is bounded and has a positive-definite covariance matrix. Analogously, define $\mathbf{S}_U = g(\mathbf{X}_{tU})$ as a function of test scores on the reference population U. An MDIA weight w_t is a positive random variable defined on study population t such that the Kullback-Leibler (Kullback & Leibler, 1951) discriminant information $E(w_t \log(w_t))$ is minimized subject to the constraints that

$$E(w_t) = 1 \tag{1}$$

and

$$E(w_t \mathbf{S}_t) = E(\mathbf{S}_U). \tag{2}$$

A unique positive constant c_t and a unique vector $\boldsymbol{\beta}_t$ exist such that w_t is an MDIA weight if, and only if,

$$w_t = c_t \exp(\boldsymbol{\beta}_t' \mathbf{S}_t) \tag{3}$$

with probability 1, and w_t satisfies Equations 1 and 2 (Berk, 1972; Csiszár, 1975; Haberman, 1984). Throughout the article, it is assumed that the MDIA weight w_t is selected according to Equation 3. To link scores from section j across forms, linking procedures can then be based on the weighted grade function $F_{jt}(x)$ of the $\mathbf{X}_{jtt}$. Let $I_x(y)$ be a comparison function such that $I_x(y) = 0$ if $x < y$, $I_x(y) = 1$ if $x > y$, and $I_x(y) = 1/2$ if $x = y$. Then, the weighted grade function for section j on administration t is given by

$$F_{jt}(x) = E(w_t I_x(X_{jtt})). \tag{4}$$

This function may be called a weighted percentile rank function (Kolen & Brennan, 2014). Thus, equipercentile linking may be applied to test scores X_{jtt} and X_{juu} by the use of $F_{jt}(x)$ and $F_{ju}(x)$ to link forms t and u.

Consider section j on form t, $1 \leq j \leq k$ and $1 \leq t \leq T$. In ideal cases in which the administrations provide equivalent groups, the weighted grade function $F_{jt}(x)$ in Equation 4 is equal to the unweighted grade function

$$F_{jtU}(x) = E(I_x(X_{jtU})) \tag{5}$$

for all x on the reference population U, so that the weighted grade function of $\mathbf{X}_{tt}$ is the same as the unweighted grade function of $\mathbf{X}_{tU}$. In such cases, $w_t = 1$ with probability 1.

A somewhat idealized case arises in the case of poststratification equating. For a positive integer $k' < k$, consider the use of k' anchor scores, so that for all test forms, k' test scores are derived from identical subtests. For convenience, suppose that the anchor scores are the first k' scores, although the order is arbitrary. Thus, one may let the anchor scores $\mathbf{A}_t$ for study population t be the k'-dimensional vector with elements X_{jtt} for $1 \leq j \leq k'$, where $X_{jtt} = X_{jtu}$ for test forms t and u. Assume that the $\mathbf{A}_t$ have values in a finite set $\mathcal{A}$ with $d + 1$ members and that $P(\mathbf{A}_t = \mathbf{a}) > 0$ for all $\mathbf{a}$ in $\mathcal{A}$. Based on Equations 4 and 5, for any x,

$$F_{jt}(x) = \sum_{\mathbf{a} \in \mathcal{A}} w_t P(\mathbf{A}_t = \mathbf{a}) E(I_x(X_{jtt}) | \mathbf{A}_t = \mathbf{a}), \tag{6}$$

and

$$F_{jtU}(x) = \sum_{\mathbf{a} \in \mathcal{A}} P(\mathbf{A}_U = \mathbf{a}) E(I_x(X_{jtU}) | \mathbf{A}_U = \mathbf{a}). \tag{7}$$

Under the assumption commonly employed in poststratification equating that, for each test form t, the conditional distribution of $\mathbf{X}_{tu}$ given $\mathbf{A}_u = \mathbf{a}$ in $\mathcal{A}$ is the same for each study population u (i.e., $\mathbf{X}_{tu}$ is population invariant), it follows that using MDIA weight w_t equal to $P(\mathbf{A}_U = \mathbf{a})/P(\mathbf{A}_t = \mathbf{a})$ whenever $\mathbf{A}_t = \mathbf{a}$ in $\mathcal{A}$ yields $F_{jt}(x) = F_{jtU}(x)$ for each test form t, every test score j, and all x. This MDIA weight is produced if, for all $\mathbf{a}$ in $\mathcal{A}$, $S_{bt} = 1$ for $\mathbf{A}_t = \mathbf{a}$ and $S_{bt} = 0$ for $\mathbf{A}_t \neq \mathbf{a}$. Test scores X_{jtt} and X_{juu}, $j > k'$, can then be linked with equipercentile methods by use of the weighted grade functions $F_{jt}(x)$ and $F_{ju}(x)$.

In the applications in this article, each administration t has $k = 4$ test scores, where $k' = 2$ equated unrounded section scaled scores X_{jtt} ($j = 1,\ 2$) are from MC sections (listening and reading), and the remaining test scores X_{jtt} ($j = 3,\ 4$) are raw scores from CR sections (speaking and writing). As noted above, the CR sections do not share common items across test administrations, so their scores are not linked. In addition, the administrations do not provide equivalent groups. Thus, cases will not be ideal because the weighted grade function $F_{jt}(x)$ is not identical to $F_{jtU}(x)$ for all x. However, rather than taking no action at all, it may well be much more effective to link scores from the CR sections using scores from the MC sections. The linking uses pseudo-equivalent groups (Haberman, 2015). This procedure is likely to be successful for assessments with high correlations between individual scores from the MC sections and individual scores from the CR sections (e.g., see Table 2 for an example). That is because statistical variability from sampling decreases as the correlations increase, a feature exploited by both MDIA (Haberman, 1984) and equating involving anchor tests (Dorans, Moses, & Eignor, 2010). A

typical recommendation is "the correlation should be as high as possible" (Dorans et al., 2010, p. 16). For assessments with large samples, the administration-level correlations between section scores are also critical (see Table 3 for an example), for they reveal the relationship between the mean section scores from the CR sections and those from the MC sections across forms. Use of multiple sections with MDIA has the added advantage that the residual variance of an individual section score or a mean section score for an administration when predicted by multiple section scores may well be substantially smaller than the corresponding residual variance achieved by prediction by a single section score. MDIA can be used much more readily with multiple section scores than can classical linking methods.

Instead of a vector $\mathbf{A}_t$ of classical anchor scores that measure the same construct as the test scores to be linked, the matching variables employed in the linking procedure are the k' equated unrounded section scaled scores X_{jtt}, $1 \leq j \leq k'$, from the MC sections that measure content areas different from the CR section scores to be linked. These scaled scores are always bounded in practice. The possible values of the X_{jtt} are not necessarily the same for all values of t. One possible choice of the d-dimensional linear constraints $\mathbf{S}_t$ to be satisfied in the MDIA procedure is based on first and second moments. One might have $d = k'(k'+3)/2$, with $S_{bt} = X_{btt}$ for $1 \leq b \leq k'$ and $S_{bt} = X_{qr}X_{rt}$ for $1 \leq q \leq r \leq k'$ and $b = k' + q + r(r-1)/2$. In this way, the weighted mean $E(w_t\mathbf{X}_{tt}) = E(\mathbf{X}_{tU})$, and the weighted covariance matrix $E(w_t[\mathbf{X}_{tt} - E(w_t\mathbf{X}_{tt})][\mathbf{X}_{tt} - E(w_t\mathbf{X}_{tt})]')$ is the covariance matrix $\mathrm{Cov}(\mathbf{X}_{tU})$ of $\mathbf{X}_{tU}$. Another possible choice of $\mathbf{S}_t$ is solely based on first moment; that is, the weighted mean $E(w_t\mathbf{X}_{tt}) = E(\mathbf{X}_{tU})$. In this case, one may have $d = k'$ with $S_{bt} = X_{btt}$ for $1 \leq b \leq k'$. This choice was considered in the data example in this article. The analysis and results described in detail in the next section are based on matching the first moments of listening scaled scores and reading scaled scores in MDIA, with $d = 2$ linear constraints and $S_{bt} = X_{btt}$ for $b = 1,\ 2$.

In practice, estimation must be employed. Let $\mathbf{X}_{tti}$, $1 \leq i \leq n_t$, $1 \leq t \leq T$, be independent observations for examinee i such that $\mathbf{X}_{tti}$ and $\mathbf{X}_{tt}$ have the same distribution. The total number of examinees n_+ is then the sum of the n_t. Let n_t/n_+ converge to a positive constant f_t as n_+ becomes large. Let f_{tU} be nonnegative random variables with sum 1 such that f_{tU} converges to $P(U = t)$ with probability 1 as n_+ becomes large. It is often the case in linking applications that $f_{tU} = n_t/n_+$ and $P(U = t) = f_t$. Let $\mathbf{S}_{ti}$ be $g(\mathbf{X}_{tti})$ for $1 \leq i \leq n_t$ and $1 \leq t \leq T$. Then, sample weights $\hat{w}_{ti}$, which are constrained to be positive, are obtained for each examinee i in form t by solving the equations

$$n_t^{-1} \sum_{i=1}^{n_t} \hat{w}_{ti} = 1, \tag{8}$$

$$n_t^{-1} \sum_{i=1}^{n_t} \hat{w}_{ti}\mathbf{S}_{ti} = \sum_{u=1}^{T} (f_{uU}/n_u) \sum_{i=1}^{n_u} \mathbf{S}_{ui}, \tag{9}$$

and

$$\hat{w}_{ti} = \hat{c}_t \exp(\hat{\boldsymbol{\beta}}_t'\mathbf{S}_{ti}). \tag{10}$$

(Haberman, 1984, 2015). The weighted grade function $F_{jt}(x)$ then has estimate $\hat{F}_{jt}(x)$ such that

$$\hat{F}_{jt}(x) = n_t^{-1} \sum_{i=1}^{n_t} \hat{w}_{ti} I_x(X_{jtti}). \tag{11}$$

Thus, $\hat{F}_{jt}(x)$ is the weighted portion of examinees i with $x > X_{jtti}$ plus half of the weighted portion of examinees with $x = X_{jtti}$.

Once the weighted sample grade function $\hat{F}_{jt}(x)$ for each form t is available, many approaches associated with randomly equivalent groups can be applied to place raw scores on different forms on a common scale. For example, once $\hat{F}_{jt}(x)$ and $\hat{F}_{ju}(x)$ are available for forms t and u on section j, many conventional equipercentile linking methods with randomly equivalent groups can be employed to equate test scores X_{jtt} and X_{juu} when they are both discrete (Kolen & Brennan, 2014; von Davier et al., 2004). Raw scores on form t (new form) may be linked to raw scores on form u (old form), and then the linked raw scores may be converted to scaled scores based on the scale for the old form. When a reference population is already defined for all forms for MDIA, scaled scores on the reference population may also exist to permit a direct conversion from a raw score on form t to a scaled score on the reference population (Haberman, 2015). This article considers such a direct conversion through weighted grade functions (Haberman, 2017). The raw scores on form t and the scaled scores on the reference population can be discrete or continuous. This is a nonparametric equipercentile equating procedure (i.e., no distribution assumptions made), and is favorable in cases with uneven distributions for raw scores on form t and for scaled scores on the reference population.

Let $G(y)$ be the unweighted grade function of a continuous random variable for scaled scores ranging from B_1 and B_2. The function $G(y)$ is a nonnegative real function such that $G(y) = 0$ if $y \leq B_1$, $G(y) = 1$ if $y \geq B_2$, and $G(y)$ is continuous and strictly increasing on the closed interval $[B_1, B_2]$. Let G^{-1} be the continuous function on [0,1] such that $G^{-1}(G(y)) = y$ for $B_1 \leq y \leq B_2$. Then, raw score X_{jtti} is converted to scaled score

$$\hat{Y}_{jtti} = G^{-1}(\hat{F}_{jt}(X_{jtti})), \tag{12}$$

which is an approximation to scaled score $Y_{jtti} = G^{-1}(F_{jt}(X_{jtti}))$ if $F_{jt}(x)$ and weight w_{ti} were both known. Because $|\hat{F}_{jt} - F_{jt}|$ converges to 0 with probability 1, the maximum difference $|\hat{Y}_{jtti} - Y_{jtti}|$, $1 \leq i \leq n_t$, converges to 0 with probability 1 (Haberman, 2015).

Data Example: Analysis and Results

As noted earlier, the reference population in the study consisted of all administrations in the year. For MDIA, $d = 2$ linear constraints $\mathbf{S}_t$ were considered, which were solely based on the first moment of listening and reading scaled scores, with $S_{bt} = X_{btt}$ for $b = 1,\ 2$. In other words, examinees in a given administration were weighted so that they had weighted equated unrounded listening and reading scores with the same means as those observed for the reference population.

Table 4
Summary Statistics of Individual Unrounded Scaled Scores

Section	Type	*N*	Mean	*SD*
Speaking	Unlinked	380,533	21.399	4.394
	Linked	380,533	21.396	4.428
Writing	Unlinked	380,533	20.025	4.825
	Linked	380,533	20.026	4.837

For the data example under study, adding the second moments of listening and reading scaled scores to the MDIA procedure with the first moments only did not have any material effect on the examinee weights or on the subsequent linked scaled scores for speaking and writing, so the rest of the discussion will focus on weights that match only the first moments and the corresponding linking results.

At the end of the MDIA step, a weight $\hat{w}_{ti}$ was estimated for each examinee in administration t, $1 \leq t \leq 42$. The weighted sample from each administration was then treated as if populations randomly equivalent groups for scoring speaking, and writing. For each administration t, the weighted sample grade functions for speaking and writing raw scores were evaluated as in Equation 11 for $j = 3,\ 4$. The unweighted grade function, $G(y)$, for the unlinked speaking scaled scores was produced via Equation 11 with a common weight for all examinees in the reference population. The unweighted grade function for the unlinked writing scaled scores was produced similarly. Then, for $j = 3,\ 4$, the raw scores X_{jtti} were linked directly to the continuous (unlinked) scaled scores in the reference population using the conversion given in Equation 12.

Summary Statistics of Scaled Scores With and Without Linking

A proper linking procedure should lead to similar overall statistics for the linked scores but reduce unexplained administration variability relative to the unlinked scores. To verify if the proposed linking procedure has this property, summary statistics of scaled scores were assessed at the individual level and at the administration level. Table 4 presents the summary statistics of individual scaled scores with and without linking. For each administration, means and standard deviations (*SD*s) of unrounded unlinked scaled scores of speaking and writing and those of unrounded linked scaled scores were examined. Table 5 shows the summary statistics of section mean and section *SD* across the 42 administrations. It is confirmed that linking had little impact on the summary statistics of individual scaled scores or on the average section statistics, but reduced variability in the section statistics across administrations.

Harmonic Regression

To evaluate the linking results, we examined whether linking improved scale stability of speaking and writing. Because test scores tend to vary by season, some form

Table 5
Summary Statistics of Section Statistics Across Administrations

Section	Statistic	Type	*N*	Mean	*SD*
Speaking	Mean	Unlinked	42	21.419	.441
		Linked	42	21.438	.374
Speaking	*SD*	Unlinked	42	4.360	.255
		Linked	42	4.400	.122
Writing	Mean	Unlinked	42	20.046	.615
		Linked	42	20.078	.444
Writing	*SD*	Unlinked	42	4.779	.249
		Linked	42	4.800	.152

of seasonal adjustment is useful when assessing variability in test scores. Harmonic regression (Lee & Haberman, 2013) is linear regression with sinusoidal functions to characterize seasonality in a time series. Additional predictors can be incorporated into harmonic regression to account for variations in test scores that reflect various types of population changes. In this study, harmonic regression analysis was applied to examine the effects of the proposed linking method on score stability. Let V_{jt} denote the value of a summary statistic of test score j, $j = 3,\ 4$, on administration t. Two summary statistics were considered for each j: one was the sample mean, and the other was the sample *SD*. A base model with no predictor is given by

$$\text{Model 0:}\ \ V_{jt} = \mu_{0,j} + e_{0,jt}, \tag{13}$$

where $e_{0,jt}$ are independent random variables with common mean 0 and variance $\sigma^2_{0,j}$, and $\mu_{0,j}$ is an unknown constant.

For each test score X_{jtt}, $j = 3,\ 4$, and each summary statistic V_{jt}, the base model was compared to harmonic regression models involving harmonic components to account for seasonality in the data and region fractions to account for changes in the regional distribution of examinees across administrations. For the assessment in study, 14 regions were defined worldwide for the examinee population based on where they took the test and their origin of birth. Let R_{tr} denote the fraction of examinees who took administration t in Region r, $1 \le r \le 14$, where $\sum_{r=1}^{14} R_{tr} = 1$. Sinusoidal functions are used to characterize seasonal patterns in V_{jt}, $1 \le t \le 42$, which typically repeat every year in educational assessments. To relate administration t to sinusoidal functions, variables associated with the time of administration are needed. Define variable d_t as the number of days elapsed since the beginning of the year at the time of administration t. The year-length variable T_t is equal to 365 for an ordinary year and equal to 366 for a leap year. For instance, an administration given on January 10, 2014 has $d_t = 10$ and $T_t = 365$. The mth harmonic component is expressed as $a_m \cos(2\pi m d_t / T_t) + b_m \sin(2\pi m d_t / T_t)$, where a_m and b_m are unknown coefficients to be estimated. After preliminary analysis, the final harmonic regression model involving $m = 2$ harmonic components to account

Table 6
Harmonic Regression Results for Different Section Statistics

Section	Statistic	Type	Model	Number of Predictors	RMSE	R^2	Adjusted R^2
Speaking	Mean	Unlinked	Model 0	0	.441	.000	.000
			Model 1	17	.291	.745	.564
		Linked	Model 0	0	.374	.000	.000
			Model 1	17	.151	.905	.837
Speaking	*SD*	Unlinked	Model 0	0	.255	.000	.000
			Model 1	17	.146	.808	.673
		Linked	Model 0	0	.122	.000	.000
			Model 1	17	.056	.877	.790
Writing	Mean	Unlinked	Model 0	0	.615	.000	.000
			Model 1	17	.343	.818	.689
		Linked	Model 0	0	.444	.000	.000
			Model 1	17	.180	.904	.836
Writing	*SD*	Unlinked	Model 0	0	.249	.000	.000
			Model 1	17	.141	.812	.679
		Linked	Model 0	0	.152	.000	.000
			Model 1	17	.064	.897	.824

Note. RMSE = root mean squared error of prediction.

for seasonality in the data and $r = 13$ region fractions to account for changes in the regional distribution of examinees across administrations is given by

$$\text{Model 1: } V_{jt} = \mu_{1,j} + \sum_{m=1}^{2}[a_m \cos(2\pi m d_t / T_t) + b_m \sin(2\pi m d_t / T_t)] + \sum_{r=1}^{13} c_r R_{tr} + e_{1,jt}, \tag{14}$$

where $e_{1,jt}$ are independent random variables with common mean 0 and variance $\sigma^2_{1,j}$, and $\mu_{1,j}$, c_r $(1 \leq r \leq 13)$, and a_m and b_m $(m = 1,\ 2)$ are unknown constants. The last region $r = 14$ was the largest subgroup and was excluded in Model 1 as the reference group for region. The overall predictive value of Models 0 and 1 was primarily assessed through the root mean squared error of prediction (RMSE), R^2, and adjusted R^2. Adjusted R^2 was considered because it combines information about prediction error with the number of parameters.

Results in Table 6 indicate that linking substantially reduced unexplained variability in means and *SD*s of speaking and writing scaled scores for the administrations studied. The RMSE of Model 0 was equal to the *SD* of the section statistic in Table 5, which represents the original variability in the section statistic without seasonal and regional adjustments. Take speaking means as an example. Means of unrounded unlinked scaled scores of speaking had an RMSE of .441 for Model 0, which was reduced to .291 with seasonal and regional adjustments in Model 1. The unexplained

Table 7
Values of the 5th, 25th, 50th, 75th, and 90th Percentiles of Unlinked Scaled Scores for Speaking and Writing

Section	P5	P25	P50	P75	P90
Speaking	14.044	19.163	21.722	24.282	26.841
Writing	10.894	17.252	20.548	23.478	26.055

variability in the linked speaking means was equal to .151 for Model 1, which is about 52% of the unexplained variability in the unlinked speaking means. It is clear that linking substantially reduced the unexplained variability in speaking means. The same observation can be made for the other section statistics.

Impact on Individual Scaled Scores

In addition to examining the impact of linking on scale stability, the impact of linking on individual speaking and writing scaled scores was evaluated. Pearson correlations of unrounded scaled scores with and without linking were above .995 for both cases. Next, for each case, the score range was divided into six groups based on the 5th, 25th, 50th, 75th, and 90th percentiles of the unlinked scores;[1] the corresponding scaled scores are presented in Table 7.

Individual's score changes from the unrounded unlinked scaled score to the unrounded linked scaled score were examined by score group. The distribution of changes in scores for each of the six score groups is depicted by box plots in Figures 1 and 2 for speaking and writing, respectively. In each figure, the left *Y*-axis gives the actual change in section score (linked−unlinked). To provide a context for interpreting the score change, the right *Y*-axis is in the scale of the *SD* for the unlinked scaled scores. For each section, the distribution of score changes varied across score groups in terms of *SD*, especially when contrasting the box plot for the lowest score group with the box plots for the other score groups. The medians (the band inside each box) and means (the diamond symbol in each box) of score changes were very close to zero for all score groups in both cases. The greatest changes were for the writing score in the lowest score group. The overall mean score changes were less than .003 in absolute value for both speaking and writing.

To further examine the influence of linking on the reported speaking and writing scaled scores, the linked and unlinked scaled scores were rounded to the nearest integer and truncated at 0 and 30. Individual's score change from the rounded unlinked scaled score to the rounded linked scaled score was evaluated by score group. Tables 8 and 9 present the frequency tables for score changes for each score group, based on unlinked scores, for speaking and writing, respectively. The results for no score change are highlighted in each table. The percentage of examinees is shown in parentheses.

Consider the speaking results in Table 8. The percentage of examinees with no score change ranged from 33.24% (score group 1) to 92.63% (score group 4) across score groups. Examinees in the lowest score group were most affected by linking;

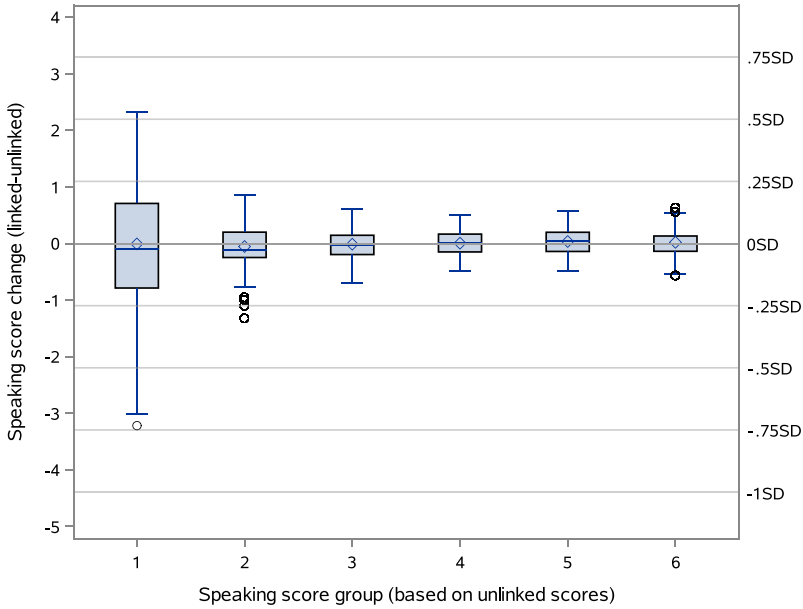

Figure 1. Box plots for each score group of changes between unrounded speaking scaled scores with and without linking. ($SD = 4.394$, see Table 4) (Color figure can be viewed at wileyonlinelibrary.com)

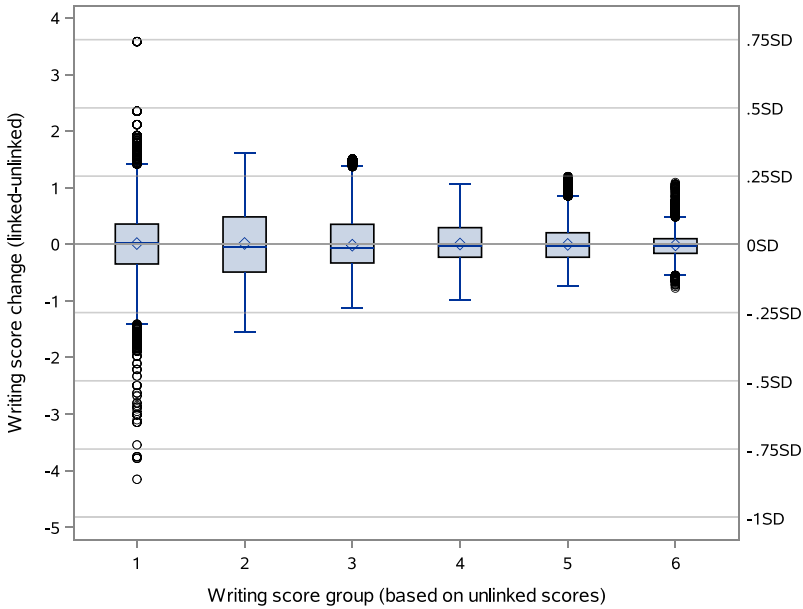

Figure 2. Box plots for each score group of changes between unrounded writing scaled scores with and without linking. ($SD = 4.825$, see Table 4) (Color figure can be viewed at wileyonlinelibrary.com)

Table 8
Frequency Table for Score Changes (From Rounded Unlinked Scaled Scores to Rounded Linked Scaled Scores) for Each Score Group: Speaking

Score Group	Speaking Score Change							
Frequency (Row Pct, %)	−3	−2	−1	0	1	2	3	Total
1	9	630	4,851	**5,259**	3,356	1,571	144	15,820
	(.06)	(3.98)	(30.66)	**(33.24)**	(21.21)	(9.93)	(.91)	(4.16)
2	0	0	17,594	**51,239**	5,656	0	0	74,489
	(0)	(0)	(23.62)	**(68.79)**	(7.59)	(0)	(0)	(19.57)
3	0	0	874	**53,603**	19,800	0	0	74,277
	(0)	(0)	(1.18)	**(72.17)**	(26.66)	(0)	(0)	(19.52)
4	0	0	7,516	**94,401**	0	0	0	101,917
	(0)	(0)	(7.37)	**(92.63)**	(0)	(0)	(0)	(26.78)
5	0	0	12,174	**44,293**	8,381	0	0	64,848
	(0)	(0)	(18.77)	**(68.3)**	(12.92)	(0)	(0)	(17.04)
6	0	0	354	**44,488**	4,340	0	0	49,182
	(0)	(0)	(.72)	**(90.46)**	(8.82)	(0)	(0)	(12.92)
Total	9	630	43,363	**293,283**	41,533	1,571	144	380,533
	(.00)	(.17)	(11.4)	**(77.07)**	(10.91)	(.41)	(.04)	(100)

Note. Row percentages are in parentheses. Cells with bold text are results for no score change.

Table 9
Frequency Table for Score Changes (From Rounded Unlinked Scaled Scores to Rounded Linked Scaled Scores) for Each Score Group: Writing

Score Group	Writing Score Change								
Frequency (Row Pct, %)	−4	−3	−2	−1	0	1	2	4	Total
1	13	33	233	3,689	**11,377**	3,594	251	21	19,211
	(.07)	(.17)	(1.21)	(19.2)	**(59.22)**	(18.71)	(1.31)	(.11)	(5.05)
2	0	0	569	18,218	**36,852**	18,768	898	0	75,305
	(0)	(0)	(.76)	(24.19)	**(48.94)**	(24.92)	(1.19)	(0)	(19.79)
3	0	0	12	17,637	**57,215**	19,107	338	0	94,309
	(0)	(0)	(.01)	(18.7)	**(60.67)**	(20.26)	(.36)	(0)	(24.78)
4	0	0	0	14,824	**66,897**	13,896	0	0	95,617
	(0)	(0)	(0)	(15.5)	**(69.96)**	(14.53)	(0)	(0)	(25.13)
5	0	0	0	8,982	**42,550**	6,670	0	0	58,202
	(0)	(0)	(0)	(15.43)	**(73.11)**	(11.46)	(0)	(0)	(15.29)
6	0	0	0	3,133	**31,549**	3,207	0	0	37,889
	(0)	(0)	(0)	(8.27)	**(83.27)**	(8.46)	(0)	(0)	(9.96)
Total	13	33	814	66,483	**246,440**	65,242	1,487	21	380,533
	(.00)	(.01)	(.21)	(17.47)	**(64.76)**	(17.14)	(.39)	(.01)	(100)

Note. Row percentages are in parentheses. Cells with bold text are results for no score change.

in particular, their score might increase by up to 3 points or decrease by up to 3 points with linking. For writing, the percentage of unaffected examinees ranged from 48.94% (score group 2) to 83.27% (score group 6) for the six score groups, and the score gain or loss could be up to 4 points with linking.

The results presented so far suggest nontrivial changes at the low end of the score scale for individual examinees. After further investigation of score conversions for low scores by administration, it was apparent that the overall distributions of reported scores at the low end are much smoother after linking. Different administrations can lead to different patterns in raw scores at the low end of the score range. To the extent that linking compensates for raw score differences due to variations in test difficulty, it should produce more consistent patterns of scaled scores at the low end. In addition, the relatively large changes at the low end reflect a standard result for sample quantile functions of continuous variables with positive density functions within their range. For any given quantile value, the variability of the corresponding sample quantile increases as the density function at the quantile decreases (Cramér, 1946, pp. 367–369). This instability, a common problem in equipercentile equating, reflects the relatively small number of observations associated with a relatively large number of low score points. Scale redefinition that reduces the number of possible scale values associated with only a small number of observations would resolve this issue.

Table 8 also shows that the distributions of speaking score changes (from rounded unlinked scaled scores to rounded linked scaled scores) were somewhat asymmetric for several score groups, but this phenomenon was not observed in the distributions of score changes for unrounded scaled scores presented in Figure 1. This asymmetry was not observed for writing with or without rounding (see Table 9 and Figure 2). One explanation is that there were gaps, or impossible reported scores, in the original speaking score scale[2] that the linking procedure filled and thereby produced a smoother distribution of scaled scores. To verify this explanation, we compared the difference between rounded and unrounded unlinked scaled scores, the difference between rounded and unrounded linked scaled scores, and the difference between these two differences for both speaking and writing for each score group. Their average changes are shown in Figures 3 and 4 for speaking and writing, respectively.

It is clear from Figure 3 that rounding did not have much impact on the linked speaking scores. However, changes in unlinked speaking scaled scores due to rounding varied considerably across score groups—ranging from .056 to .264 in absolute value. These findings are a result of rounding to adjacent scores of the gaps without linking. Compared to speaking, the effects of rounding on linked and unlinked writing scaled scores were negligible (Figure 4). It is worth noting that there were no gaps in the original writing score scale.

Comparability of Linking Results Across Regions

It is desirable that linking results be comparable across important subpopulations within the reference population. Recall that the harmonic regression analysis identified the 14 regions defined worldwide for the examinee population as important

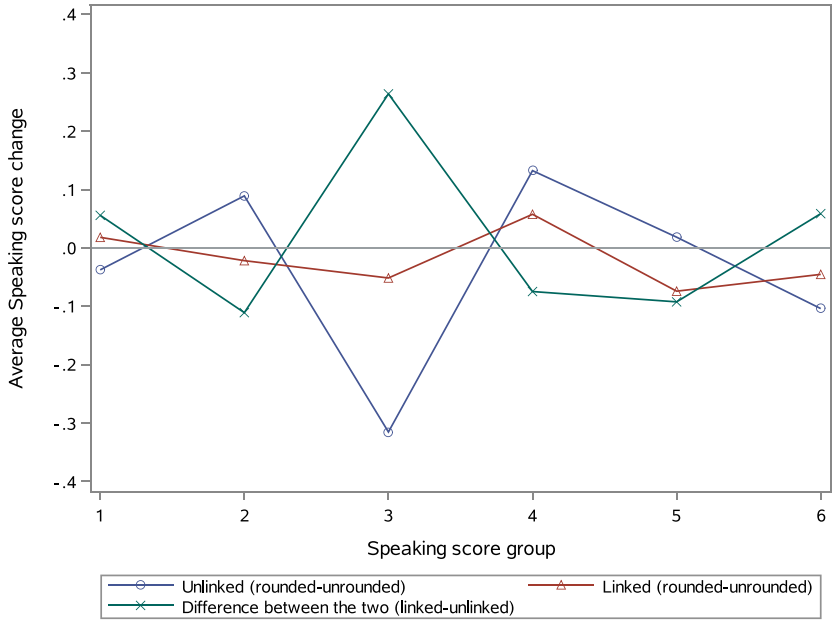

Figure 3. Average changes in speaking scaled scores with and without rounding and with and without linking. (Color figure can be viewed at wileyonlinelibrary.com)

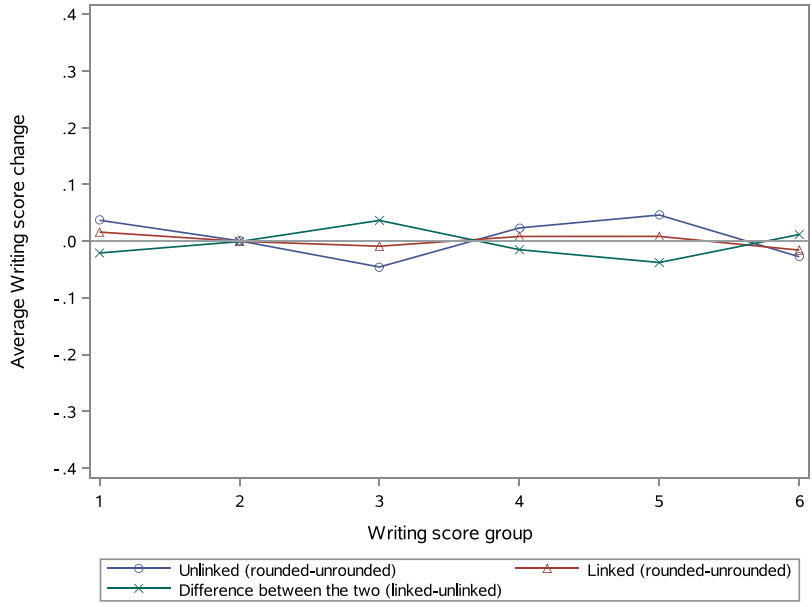

Figure 4. Average changes in writing scaled scores with and without rounding and with and without linking. (Color figure can be viewed at wileyonlinelibrary.com)

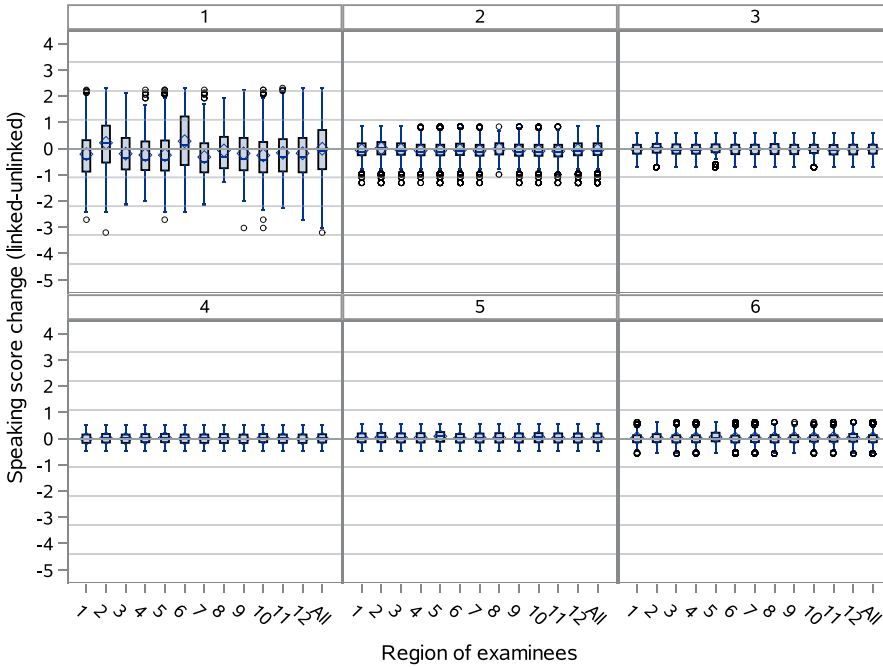

Figure 5. Box plots for each score group of changes between unrounded speaking scaled scores with and without linking, by region of examinees and for the combined group, labeled "All." (*Note.* In each subfigure, the gray horizontal reference lines from top to bottom refer to .75*SD*, .5*SD*, .25*SD*, 0*SD*, −.25*SD*, −.5*SD*, −.75*SD*, and −1*SD*, with $SD = 4.394$ from Table 4.) (Color figure can be viewed at wileyonlinelibrary.com)

subpopulations that affected speaking and writing scores. Only 12 of the 14 regions were used to assess comparability of the linking results across subpopulations for two reasons. First, one region involved merely 173 examinees (about .05% of the overall population), with score group sizes ranging from 4 (score group 1) to 52 (score group 4) for speaking and from 8 (score group 1) to 51 (score group 3) for writing. For this region, some of the score groups were too small to yield meaningful linking results. Second, the assessment in study was available throughout the year to all regions but one. This region was added to the test population late in the study year and was only exposed to 5 out of the 42 administrations. The sample size of this region was 10,591 (about 2.78% of the overall population). Because the linking procedure adjusted scores by administration, there was clear impact of administration selection on the linking of this region.[3] Thus, these two regions were excluded from the subpopulation evaluation. The box plots depicted in Figures 1 and 2 were disaggregated by region of examinees for each of the six score groups, and Figures 5 and 6 show the disaggregated results for speaking and writing, respectively. Each figure consists of six subfigures, one for each score group; in each subfigure, the *Y*-axis gives the actual change in section score (linked−unlinked) and the *X*-axis labels the regions. As was the case with Figures 1 and 2, the gray horizontal reference lines in each subfigure of Figures 5 and 6 indicate the scale of the *SD* for the unlinked scores.

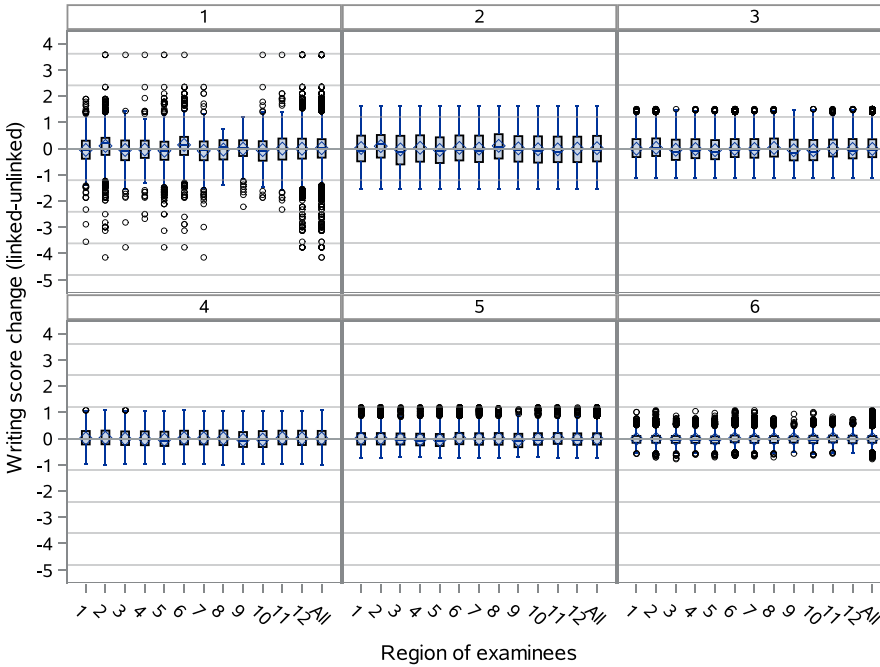

Figure 6. Box plots for each score group of changes between unrounded writing scaled scores with and without linking, by region of examinees and for the combined group, labeled "All." (Note. In each subfigure, the gray horizontal reference lines from top to bottom refer to .75*SD*, .5*SD*, .25*SD*, 0*SD*, −.25*SD*, −.5*SD*, −.75*SD*, and −1*SD*, with $SD = 4.825$ from Table 4.) (Color figure can be viewed at wileyonlinelibrary.com)

For ease of comparison between the distributions of score changes with and without disaggregation, the box plots in Figures 1 and 2 were added to and labeled as "All" in each subfigure of Figures 5 and 6. For all but the lowest score group of either speaking or writing, the distributions of score changes were very similar across the 12 regions and to the combined group. For score group 1 of each section, region 8 had less variable score changes compared to the other 11 regions and the combined group, mainly due to much smaller group sizes (43 for speaking and 129 for writing). Thus, the effects of the linking on speaking and writing scores were deemed comparable for most score groups among those important subpopulations within the reference population defined in the study.

Discussion

The purpose of linking is to adjust for the inevitable differences in difficulty that exist among test editions built to the same set of specifications. When multiple section scores of a test are highly correlated but from different content areas, it is possible to utilize the relationship between section scores to perform linking for the unlinked sections. This article presents a procedure for linking CR section scores of an assessment for which common items are unavailable based on observed equated

scores from MC sections of the test that measure contents different from the CR sections. It addresses the long-standing issue in educational measurement that linking of CR test scores in the absence of common items is impossible in many practical circumstances. With an appropriate reference population, MDIA can be used to produce pseudo-equivalent groups with respect to scores from the MC sections of the test for each administration. Then, a nonparametric equipercentile equating procedure based on weighted grade functions can be employed to adjust the CR scores by administration with a direct raw-to-scale conversion.

For the assessment under study, the use of pseudo-equivalent groups made it possible to link speaking (or writing) scaled scores across administrations in the absence of common items, something that has not been done before. Application of this linking procedure placed the scaled scores of speaking (or writing) on the same reference scale across administrations and hence the scores became more comparable. The results of this investigation determined that the proposed linking approach did reduce the variability across administrations in means and standard deviations for both speaking and writing. At the same time, individual's reported scaled scores were barely influenced for the majority of examinees. Moreover, gaps in the original speaking score scale diminished after linking and the resulting new score distributions were much smoother. Also, the results of the linkings were deemed comparable across the important subpopulations within the reference population defined in the study.

The proposed linking procedure is applicable to assessments with a structure similar to Table 1. It is assumed that the CR test to be linked is administered together with one or more MC tests with equated scores that measure different constructs than the CR test. In other words, separate scores are reported for the CR test and the MC test(s). In addition, the CR test does not have common items across test administrations. Because CR tests delivered in such a structure cannot be linked via common items, employing the procedure presented herein will enhance the comparability of the resulting CR scores from different forms. For CR tests that are part of a mixed-format test, their composite scores can be linked via existing procedures for mixed-format tests (e.g., Dorans, 2003; Kim et al., 2010; Tate, 2000).

It is important to recognize that linking can be population-dependent. In this study, the reference population was composed of all administrations of the assessment in 1 year. Although the individual administrations showed seasonality in test performance, the movements were consistent for all sections, and the correlations of mean administration section scores in Table 3 were a bit higher than the corresponding correlations of individual section scores in Table 2. In addition, the important subpopulations within this reference population had similar score profiles among the four sections in the test.

In general, it is desirable to select a reference population that is representative of the composite of the individual administrations to be linked. For applications to linking CR tests, empirical data from the most recent years may be used to construct a representative reference population for new administrations. However, if one selects a reference population that is very different from the administrations to be linked, especially in terms of performance level or score profile, the procedure may not work well and a number of findings can be expected. First, the individual's reported scores

would be influenced more than was observed in this article. Second, the distribution of score changes may show undesirable patterns across score groups and not center around 0. Third, the impact of linking on individual scaled scores may differ noticeably for some important subpopulations within the selected reference population. These expectations may be further explored in future studies. As mentioned in the subsection on subpopulation evaluation, the impact of administration selection should be examined with more simulated or real data to better understand the properties of the linking procedure. In addition, as noted above, the proposed linking procedure makes use of the relationship between section scores to perform linking for the unlinked sections. More research is needed to investigate how the strength of the correlations between sections, in the selected reference population and in the individual administrations to be linked, affects the linking procedure.

Acknowledgments

The authors would like to thank the associate editor and the three anonymous reviewers for their helpful comments. The authors also thank Hongwen Guo, J. R. Lockwood, and Rebecca Zwick for their helpful comments on an earlier version. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

Notes

[1]Score group 1: lowest 5% of scores; score group 2: $>$ 5 to 25%; score group 3: $>$ 25 to 50%; score group 4: $>$ 50 to 75%; score group 5: $>$ 75 to 90%; score group 6: $>$ 90%.

[2]There are six gaps in the original speaking score scale: 2, 7, 12, 16, 21, and 25.

[3]This region resembled region 8 in Figures 5 and 6 in terms of examinee background, proficiency level, and sample size in the data set. The major difference between them was that all 42 administrations were available to region 8, which led to balanced effects of linking for region 8 across administrations but not for the other region. As noted in the Discussion section, more research is needed to shed light on the impact of restricting when examinees can participate in an administration.

References

Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Annals of Mathematical Statistics*, *43*, 193–204.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Csiszár, I. (1975). *I* -divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*, 146–158.

DeMauro, G. E. (1992). *An investigation of the appropriateness of the TOEFL test as a matching variable to equate TWE topics* (Research Report No. RR-92-26). Princeton, NJ: Educational Testing Service.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, 423–444.

Dorans, N. J. (Ed.). (2003). *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (Research Report No. RR-03-27). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (Eds.). (2010). *Principles and practices of test score equating* (Research Report No. RR-10-29). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12*, 971–988.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*, 254–273.

Haberman, S. J. (2017). *A program for nonparametric equivalent-group equating* (Research Memorandum No. RM-17-01). Princeton, NJ: Educational Testing Service.

Hanson, B. A. (1993, April). *A missing data approach to adjusting writing sample scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, *47*, 36–53.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practice* (3rd ed.). New York, NY: Springer.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*, 815–829.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, *37*, 329–346.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.

Authors

YI-HSUAN LEE is Principal Research Scientist, Educational Testing Service, 660 Rosedale Road, MS 12T, Princeton, NJ 08541; ylee@ets.org. Her primary research interests include analysis of timing and process data, test security, quality control of assessment, item response theory, and equating and linking.

SHELBY J. HABERMAN is an independent consultant, Barak 3/1, Jerusalem 9350276, Israel; haberman.statistics@gmail.com. His primary research interests include analysis of qualitative data, sample weighting, item-response theory, equating and linking, analysis of sub-scores, and test security.

NEIL J. DORANS is Distinguished Presidential Appointee, Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541; ndorans@ets.org. His primary research interests include score linking and fairness assessment.