Measuring the Power of Learning.™

Research Report
ETS RR–18-02

The Pseudo-Equivalent Groups Approach as an Alternative to Common-Item Equating

Sooyeon Kim

Ru Lu

December 2018

ETS Research Report Series

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

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

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Pseudo-Equivalent Groups Approach as an Alternative to Common-Item Equating

Sooyeon Kim & Ru Lu

Educational Testing Service, Princeton, NJ

The purpose of this study was to evaluate the effectiveness of linking test scores by using test takers' background data to form pseudo-equivalent groups (PEG) of test takers. Using 4 operational test forms that each included 100 items and were taken by more than 30,000 test takers, we created 2 half-length research forms that had either 20 (strong anchor) or 10 (weak anchor) items in common. Because the 2 research forms were assembled from a single form that had been administered in a large-scale operational testing setting, we obtained the direct equating function between the 2 research forms through the single-group design and treated it as a criterion or true equating function between the 2 research forms. We equated the 2 research forms in a common-item design using the poststratification equipercentile (PSE) and chained equipercentile (CHEQ) methods, and then compared the common-item results to the results derived from the PEG linking. Because the new and reference groups differed substantially in ability, by study design, the CHEQ method produced more accurate results than did the PSE method in both the strong and weak anchor conditions. CHEQ using 10 common items was as effective as PSE using 20 common items. PSE using 10 common items produced the least accurate results among the five methods. The PEG linking produced more accurate results compared to the PSE method using the weak anchor.

Keywords Pseudo-equivalent groups; common-item design; poststratification equipercentile; chained equipercentile; equating

doi:10.1002/ets2.12195

A common-item equating design is often used to adjust for differences in difficulty among alternate forms of a test. Due to security concerns, however, some testing programs are unable to use items in more than one form or test administration. That is because in some test-taker populations, exposure of test items on the Internet has caused the difficulty of the items to change when they are reused. As a consequence, the most popular equating design, a common-item design (often called the nonequivalent groups with anchor test, or NEAT), cannot be implemented. It is therefore important to find an alternative design for equating new forms administered under this circumstance.

Liao and Livingston (2012) presented three approaches that could be considered as alternatives to a common-item equating design. In their paper, the randomly equivalent forms approach assembles test forms of equal difficulty by stratified random sampling of items from the item pool. The demographically adjusted group (DAG) procedure, a post-stratification procedure, uses demographic variables to create statistically matched groups of test takers. The repeaters scaling approach uses common test takers (i.e., repeaters) instead of common items to link scores on different test forms. Liao and Livingston investigated the effectiveness of these three alternatives using a large-scale international English Listening and Reading proficiency test. Because the results derived from those approaches were not promising, none of the approaches was recommended as a substitute for common-item equating for the test used in their investigation.

A Pseudo-Equivalent Groups Approach

Haberman (2015) provided a comprehensive theoretical review and procedures for pseudo-equivalent groups (PEG) linking. In general, the PEG approach uses test takers' background information (i.e., age, gender, education, major of highest education, job type, purpose for taking the test, number of times test was taken previously) to adjust for group differences in ability. Although it may sound similar to the DAG approach, the methodological model associated with PEG is more sophisticated. The PEG approach is intended to create a new-form group and a reference group (often called a target group) that are equal in ability by weighting the individual test takers in the new-form group. The weights are determined by minimizing a discriminant function formed from background variables.[1] When the relevant background differences

Corresponding author: S. Kim, E-mail: skim@ets.org

between the groups are eliminated, the scores on the two forms can be directly linked, as in an equivalent-groups equating design.

In general, PEG linking consists of three major steps. The first step is to create the target background distribution as a method to form PEGs. Assume that there are two test forms, X and Y, and each form includes the same questionnaire designed to collect test takers' background information. Let x_i be the score of Test Taker i, who received Form X, and y_j be the score of Test Taker j, who received Form Y. Let z_{iX} and z_{jY}represent the K-dimensional Vector Z of background variables collected through the questionnaire for Test Taker i on Form X and for Test Taker j on Form Y, respectively.[2] Under this condition, the target background mean is the averaged background Z vectors over the forms, as shown in Equation (1).

$$\overline{Z} = \left(\sum_{j=1}^{N_X} z_{jX} + \sum_{i=1}^{N_Y} z_{iY} \right) / \left(N_X + N_Y \right), \tag{1}$$

where N_X and N_Y are the number of test takers on Form X and Form Y, respectively. $\overline{Z}$ is then used as the target background vector to construct PEGs. In this step, each test taker on Form X gets a weight w_{iX} so that the weighted background vector of the new-form group is matched to the target Vector $\overline{Z}$, as shown in Equation (2).

$$\sum_{i=1}^{N_X} w_{ix} z_{ix} / N_X = \overline{Z}, \tag{2}$$

where $w_{iX} > 0$ and $\sum_{i=1}^{N_X} w_{ix} = 1$. Here the adjustment of the minimum discriminant information approach with the Newton – Raphson method is used to obtain the individual weight w_{iX} (Haberman, 2015). There is no limit in terms of the number of separate test administrations that can be used for the target background vector construction. For example, Haberman in this empirical investigation used the weighted average background Vector $\overline{Z}$ derived across 29 operational administrations to represent the target background distribution.

The second step is to obtain a target score distribution and a new-form score distribution. In the example of Haberman (2015), the target score distribution was defined by pooling the 29 operational scaled scores (i.e., equated) together. The target score distribution can also be defined using various functions, such as the exponential family (Haberman, 2010). The same score function should then be applied for the new form. The essence of the PEG linking is that the weighted raw scores are used to define the new-form score distribution instead of the actual raw scores on Form X to eliminate the relevant differences between the groups.

The last step is to conduct score linking using the two score distributions from the previous step: Link the raw score distribution of the weighted new-form group (i.e., the weighted raw score distribution) to the target score distribution. Because the new-form group is considered to be pseudo-equivalent to the target (reference) group after applying the weight to each new-form test taker, the PEG linking can use any conventional equivalent group equating method such as mean – sigma (linear) or equipercentile (nonlinear).

The PEG approach has been examined in a number of studies, not only where anchors are unavailable but also where effective anchor tests are available. Haberman (2015) examined a series of test forms that were equated through a common anchor test in an operational setting to demonstrate the effectiveness of the PEG approach. The PEG linking results were similar, but not identical, to common-item equating. The overall score distributions derived from the PEG approach were fairly comparable with the results from the conventional anchor equating. Haberman mentioned that if linking is based on PEG for a testing program expected to last for many years, it would be appropriate to seek some verification that the relationship of background variables to test takers' performance remains stable over time.

Recently, several studies were conducted to assess the practical implications of PEG linking for large-scale assessment programs. Oh, Liu, and Gaj (2015) applied PEG linking as well as multiple linear regression to a large-scale K – 12 assessment to see which method would be more effective in adjusting testing mode differences of online and paper versions of this assessment in a situation where previous scaled scores were not available and groups were not equivalent. The researchers compared the resulting conversions from the two methods to a criterion conversion, which was the paper-mode conversion, using several deviance measures. In their comparison, the PEG linking approach was not as effective as the regression approach, particularly at the extreme score regions, leading to slightly larger overall mean squared errors. Xi, Guo, and Oh (2015) compared PEG linking to conventional linking through anchor items using the data from a large-scale

Table 1 Descriptive Statistics for the Four Pairs of Research Forms to be Equated Calculated From Total Group Data

			New Form X		Reference Form Y		
Operational form	Items in each research form	Number of test takers	Mean	*SD*	Mean	*SD*	*SMD* (X − Y)
1	60	109,754	45.53	8.53	47.32	8.23	−.21
2	60	33,067	40.14	9.77	37.50	9.78	+.27
3	60	113,787	40.10	9.39	38.17	9.32	+.21
4	60	68,951	32.45	10.10	35.01	10.41	−.25

Note. SMD = standardized mean difference in difficulty between new Form X and reference Form Y in the total group. It is worth noting that the conventional use of *SMD* is to compare the group difference in ability. In this case, however, we used the *SMD* to compare the form difference in difficulty.

standardized test. The two new forms (X1 and X2) of the test did not share any items in common, but they shared some items in common with the reference form (Y). In practice, the chained equipercentile (CHEQ) method in a common-item design was used to produce the operational conversion. For the PEG linking, they used both background variables (i.e., gender, region, grade, ethnicity, and first and best language) and the scores on the anchor items as the matching variables and compared the difference between the PEG conversion and the operational one. PEG linking through only background variables performed poorly due to the limited background information. However, when the PEG linking was conducted using not only the background variables but also the scores on the anchor items, the difference between PEG and the conventional linking disappeared. Under these conditions, the PEG approach performed as well as its counterpart. The anchor score was a single effective matching variable in the weighting procedure for minimizing discriminant information between the two groups.

When it comes to test score equating, the use of real data has the limitation that the true relationship between forms is unknown. Using simulated datasets, Lu and Guo (2015) examined the effectiveness of the PEG linking approach under both random equivalent groups and common-item (i.e., NEAT) designs in a situation where a true criterion existed. They compared three equating/linking methods: PEG, NEAT, and a hybrid of PEG and common-item (which the authors named PEG-EAT) under various testing situations manipulated by several factors such as group ability difference, anchor length, and the relationship between background variables and test scores. The PEG approach performed better than the NEAT approach only when the anchor set was unsatisfactory.

The purpose of this study is to compare the PEG approach to common-item equating (i.e., CHEQ and PSE) in terms of accuracy for equating test forms taken by groups that differ in ability. These comparisons are made by simulating a situation where the true equating relationship of two test forms in a test-taker population is known. Therefore, the accuracies of the common-item equating and from PEG linking were assessed by directly comparing them to the true equating function.

Method

Data

We chose four operational forms that included 100 multiple-choice items each and had been taken by more than 30,000 test takers in a single administration. Using each of the four selected operational test forms as an item pool, we created a pair of research forms from each operational form for the study. This study involved the equating of four pairs of test forms through common items. Therefore, the two research forms in each pair shared either 20 (long/strong anchor) or 10 (short/weak anchor) items in common. The two forms in each pair were equal in length (60% as long as the parent form) and were built to be parallel in content but unequal in difficulty. Table 1 shows a statistical comparison of the two research forms created from each of the four operational forms. In two of the four pairs of research forms, the new form was more difficult than the reference form; in the other two pairs, the new form was easier than the reference form. As the standardized mean difference (*SMD*) indicates, the size of the mean difference in difficulty between the two research forms varied from 0.21 to 0.27 in standard deviation units.[3]

After constructing a pair of research forms from each of the four operational forms, we divided the test takers who took the test at the same administration into two groups based on their previous experience with the test. Historically, repeaters of this assessment tend to perform better than do first-timers. When dividing one group into two, we assigned more repeaters to one group (will be more able) than the other (will be less able) to make them dissimilar in ability, thus

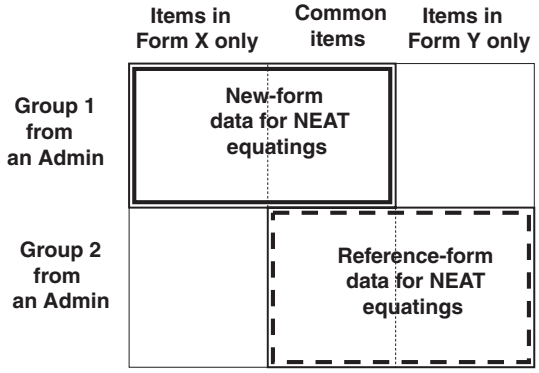

Figure 1 Design for the nonequivalent groups with anchor test design.

Table 2 Test Form Difference in Difficulty and Test-Taker Group Difference in Ability

Dataset case	New form	New group	Reference form	Reference group
1A	Hard	Strong	Easy	Weak
1B	Hard	Weak	Easy	Strong
2A	Easy	Weak	Hard	Strong
2B	Easy	Strong	Hard	Weak
3A	Easy	Strong	Hard	Weak
3B	Easy	Weak	Hard	Strong
4A	Hard	Weak	Easy	Strong
4B	Hard	Strong	Easy	Weak

creating a nonequivalent groups design. Figure 1 illustrates the NEAT design framework used in this study. We used the more difficult form as the new form in two cases and as the reference form in the other two cases. Because both test-taker groups actually took all the items in both research forms, we also had to decide which group to use as the new-form group and which as the reference form group. We decided to create two different datasets for each pair of research forms, using the more able group as the new-form group in one dataset and as the reference form group in the other dataset.

Table 2 shows a layout of the two research forms and of the two test-taker populations in each of the eight datasets used for the comparison. The two research forms in Dataset 1A and Dataset 1B are the same. The two test-taker groups in Dataset 1B are exactly reversed from those in Dataset 1A. This pattern is repeated in the other three pairs of datasets. Given that the direction of the equating should not matter (because equating is symmetric), those eight comparisons which can be classified into two cases. One is the case where the stronger group takes the more difficult form (1A, 2A, 3B, and 4B), and the other is the case in which the stronger group takes the easier form (1B, 2B, 3A, and 4A). Each case includes four replications, two in which the more difficult form is the new form and two in which the more difficult form is the reference form.

Table 3 presents the means and standard deviations of the raw (number-correct) scores of the total and anchor tests in each of the new and reference (old) form groups. The correlations between the total scores and anchor scores are also included. As the anchor scores indicate, the test takers on one form tended to differ systematically in ability from those on the other form, by design. The size of the difference between the anchor means of the new and reference form groups varied from 0.16 to 0.20 in standard deviation units.[4] The correlations between the total score and anchor score ranged from 0.87 to 0.92 with the 20 common items, but from .74 to .84 with the 10 common items.

Criterion

For each pair of forms equated in this study, the criterion equating was calculated by using the total number of test takers in the operational setting. Using the scores of all test takers from a single administration, we performed a direct equipercentile equating of scores on the new form to scores on the reference form in the total group to obtain the criterion equation function. In Figure 2, four plots display the differences between the criterion function and the identity function (i.e., no

Table 3 Descriptive Statistics for the Total and Anchor Scores in Both New and Reference Form Groups

		Total		Long anchor			Short anchor		
Data	Statistic	*X*	*Y*	vx_20	vy_20	*SMD*	vx_10	vy_10	*SMD*
1A	*N*	54,877	54,877	–	–		–	–	
	M	46.26	46.58	15.63	15.15	0.16	7.74	7.51	0.13
	SD	8.04	8.68	2.89	3.17		1.64	1.77	
	r	–	–	0.88	0.90		0.74	0.77	
1B	*N*	54,877	54,877	–	–		–	–	
	M	44.79	48.05	15.15	15.63	−0.16	7.51	7.74	−0.13
	SD	8.92	7.68	3.17	2.89		1.77	1.64	
	r	–	–	0.90	0.89		0.77	0.75	
2A	*N*	16,533	16,534	–	–		–	–	
	M	39.07	38.52	12.36	13.11	−0.20	5.98	6.42	−0.19
	SD	10.04	9.42	3.96	3.70		2.37	2.26	
	r	–	–	0.92	0.91		0.84	0.83	
2B	*N*	16,534	16,533	–	–		–	–	
	M	41.21	36.47	13.11	12.36	0.20	6.42	5.98	0.19
	SD	9.37	10.03	3.70	3.96		2.26	2.37	
	r	–	–	0.91	0.92		0.83	0.84	
3A	*N*	56,894	56,893	–	–		–	–	
	M	41.22	37.22	13.24	12.69	0.17	5.62	5.31	0.16
	SD	8.63	9.80	2.97	3.34		1.89	1.99	
	r	–	–	0.87	0.90		0.74	0.78	
3B	*N*	56,893	56,894	–	–		–	–	
	M	38.99	39.11	12.69	13.24	−0.17	5.31	5.62	−0.16
	SD	9.96	8.70	3.34	2.97		1.99	1.89	
	r	–	–	0.90	0.88		0.77	0.75	
4A	*N*	34,476	34,475	–	–		–	–	
	M	31.38	36.10	10.85	11.62	−0.20	4.64	4.95	−0.15
	SD	10.22	10.12	3.85	3.71		2.02	2.02	
	r	–	–	0.91	0.91		0.76	0.74	
4B	*N*	34,475	34,476	–	–		–	–	
	M	33.52	33.92	11.62	10.85	0.20	4.95	4.64	0.15
	SD	9.86	10.60	3.71	3.85		2.02	2.02	
	r	–	–	0.90	0.91		0.75	0.75	

Note. *N* = number of test takers; vx_20 = the anchor score based on the 20 common items in the Form X group; vx_10 = the anchor score based on the 10 common items in the Form X group. Both vy_20 and vy_10 indicate the same type of anchor scores in the Form Y group. *r* indicates the correlation between the total score and the anchor score; *SMD* = standardized mean difference between the new group and the reference group in ability based on the anchor scores.

equating) derived from each of the four datasets used in this study. Because the identity function is the function for forms that are completely parallel, the difference between the two functions indicates the extent to which equating is necessary due to the form difference in difficulty.

Procedure

In this study, we compared five linking methods. They are (a) poststratification equipercentile (PSE, often called frequency estimation equipercentile; Kolen & Brennan, 2004, pp. 135–143) using 20 common items (PSE 20); (b) PSE using 10 common items (PSE 10); (c) CHEQ (Kolen & Brennan, 2004, pp. 145–147) using 20 common items (CHEQ 20); (d) CHEQ using 10 common items (CHEQ 10); and (e) PEG.[5]

Using the PSE and CHEQ methods, we equated the new form to the reference form using either 20 or 10 common items in the NEAT design framework illustrated in Figure 1. Then we computed the difference between the equating function and the criterion equating function at each new-form raw score level. Computing the difference at each new-form raw score level is important in comparing linking methods because a linking method can be accurate in some score regions but inaccurate in others. Furthermore, we computed the mean and standard deviation (*SD*) of the equated raw scores by

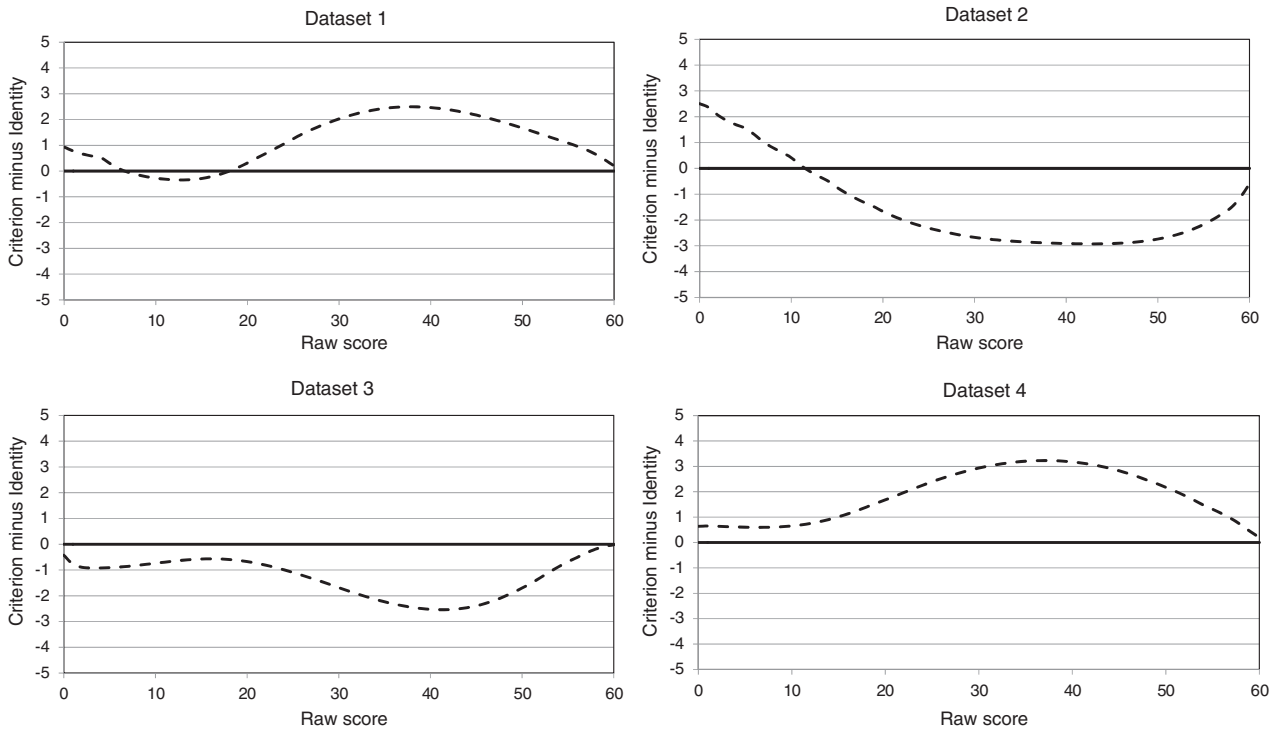

Figure 2 The difference plots between the direct equipercentile criterion function and the identity function in each of the four datasets used in this study.

applying each conversion to the test takers on the new-form group and then comparing the resulting equated summary statistics to the criterion mean and *SD*.

The PEG linking was applied to each new-form group using the test takers' background information rather than their anchor scores. In practice, the distributions of background variables should be derived from a substantial amount of data to best represent the target population taking the test. In this study, we used the entire group of test takers (new plus reference) to define the target background distribution to mimic the actual PEG linking situation. We used the 15 background variables to form a target weight vector to construct the pseudo population. The weight for each test taker in the entire group was computed by the minimum discriminant information approach (Haberman, 2014). Then the weights were applied to the raw scores of the new-form sample to obtain the weighted raw score distribution. The weighted raw score distribution on the new-form group was linked to the raw score distribution of the target (reference) group using the equipercentile method (Kolen & Brennan, 2004, pp. 36–48). Using the conversion derived from the PEG linking, we calculated the difference from the criterion.

The differences among the conversions were also quantified using the root mean squared difference (*RMSD*),

$$RMSD = \sqrt{\sum_{i=0}^{60} w_i \left[\widehat{e}_i\left(x_i\right) - e_i\left(x_i\right)\right]^2}, \tag{3}$$

where i represents a raw score point, $\widehat{e}_i\left(x_i\right)$ is the equated scores of a linking method at raw score x, $e_i(x_i)$ is the criterion equating function at raw score x, and w_i is the relative proportion of the new-form test takers at each score point. To display the score region where most test takers are located, the raw score distributions of the new-form group are presented in Figure 3. The four plots in Figure 3 present the relative frequency distributions of the new-form scores in each of the four datasets, respectively. In each plot, one distribution is associated with the new-form group in Case A, and another is associated with the new-form group in Case B.

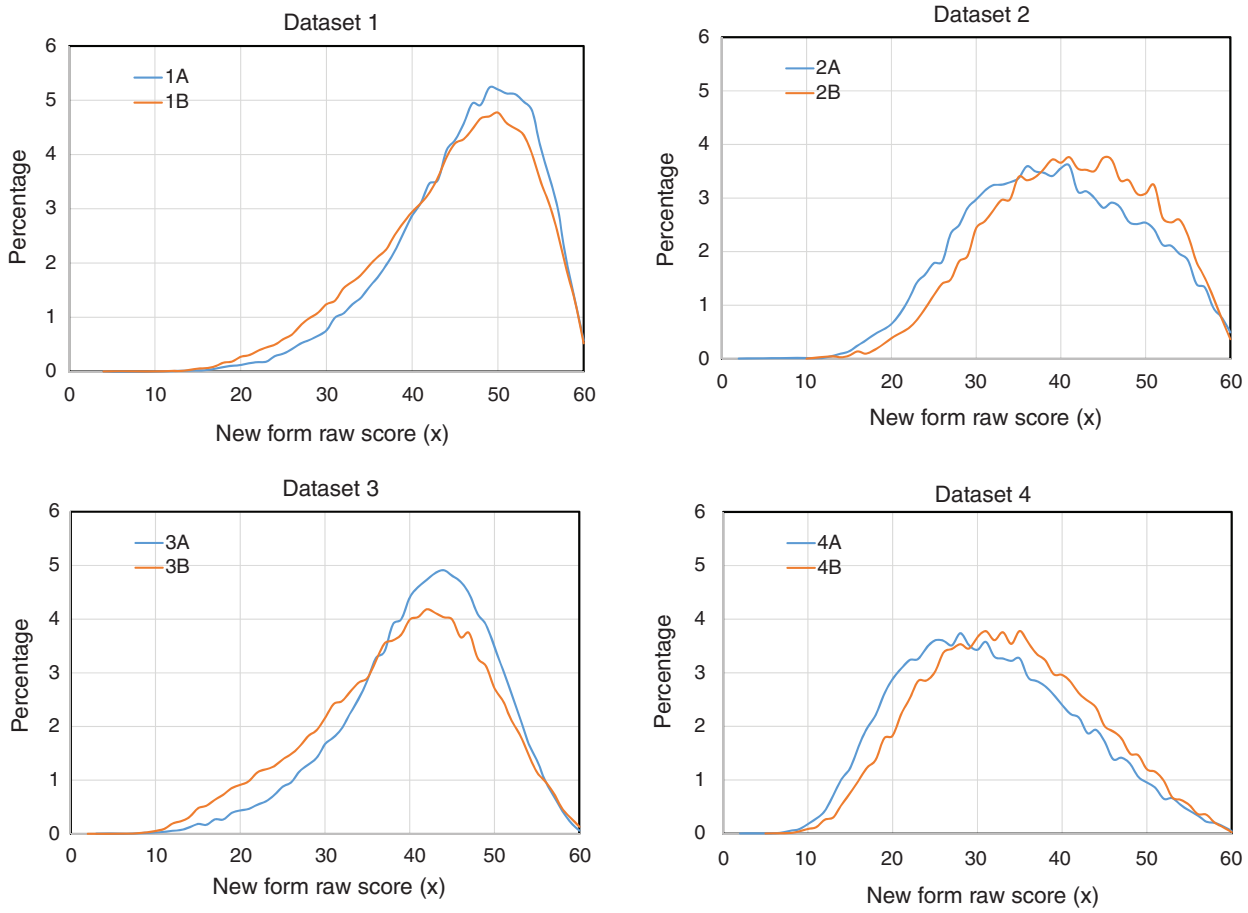

Figure 3 The raw score distribution of the new-form group in each of the two cases (A and B) in the four datasets used in this study.

Results

Figures 4 – 11 plot the differences from the criterion derived from the five linking methods across the raw score region from the 1st percentile to the 99th percentile in the new-form group. The dotted lines at ±0.5 indicate the *difference that matters* (DTM; Dorans & Feigenbaum, 1994), defined as half a raw score point. In each figure, the solid and dotted lines in blue indicate the difference derived from PSE 20 and PSE 10, respectively. The solid and dotted lines in red indicate the difference derived from CHEQ 20 and CHEQ 10, respectively. The solid black line indicates the difference associated with the PEG linking. Table 4 presents the *RMSD* values, summarized over the score distribution of the new-form group, for each of the five linking methods. Table 5 presents the means and *SD*s of the equated (or linked) raw scores associated with each of the five linking functions in the new-form group, along with the mean and *SD* derived from the criterion function.

Figures 4 and 5 present the differences from Dataset 1A and Dataset 1B, respectively. Because the new and reference groups in Dataset 1A were swapped in the Dataset 1B case, the negative differences in Figure 4 became positive differences in Figure 5. In both the 1A and 1B cases, the CHEQ method performed better than did the PSE method when they were compared under the same anchor condition. As expected, the CHEQ 20 method produced the smallest difference across the score region where most test takers were located. The CHEQ 10 method was as effective as the PSE 20 method. The differences associated with CHEQ 20 were generally within the DTM band. The differences associated with PSE 20 were slightly larger than the DTM across the raw score region from 24 to 38. As shown in Table 4, CHEQ 20 led to the smallest *RMSD*s among the five methods, and its magnitude was smaller than 0.5 in both cases. Both CHEQ 10 and PSE 20 also led to *RMSD*s smaller than 0.5. Accordingly, as presented in Table 5, the means and *SD*s of the equated raw scores associated with those methods were fairly comparable to the mean/*SD* of the criterion conversion. In the Dataset 1A case, the magnitudes of differences associated with PEG and PSE 10 were similar for the score region above 30, but PSE 10

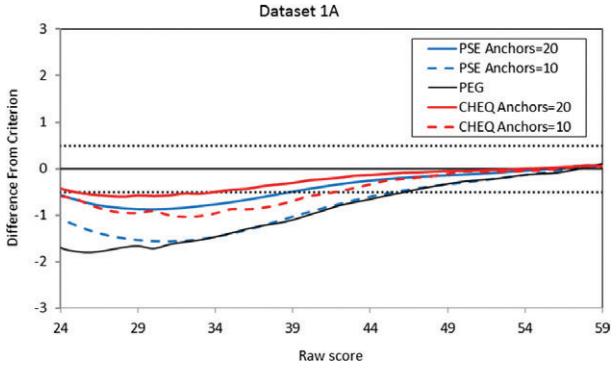

Figure 4 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 1A case.

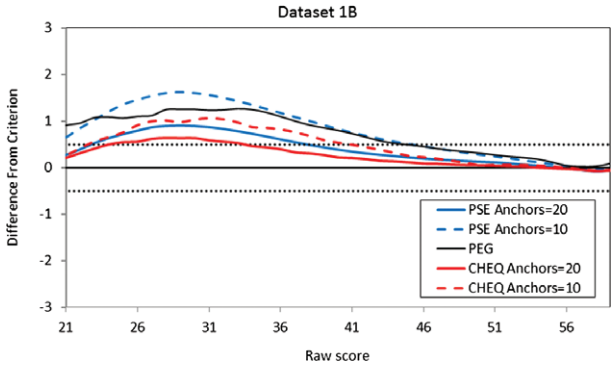

Figure 5 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 1B case.

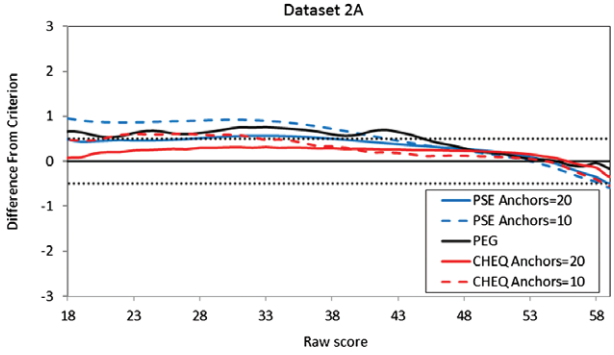

Figure 6 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 2A case.

outperformed for the score region below 30. This trend was somewhat opposite in Dataset 1B. PEG performed better than did PSE 10 for the score region below 35. The magnitudes of overall *RMSD*s associated with the two methods were larger than 0.5 in both the 1A and 1B cases. Regardless of the direction (either positive or negative), the absolute mean differences of the PSE 10 and PEG methods from the criterion mean were greater than 0.5 in both cases.

Figures 6 and 7 present the differences from Dataset 2A and Dataset 2B, respectively. Because the new and reference groups in the 2A case were swapped in the 2B case, the positive differences in Figure 6 became negative differences in Figure 7. In both cases, the CHEQ 20 method produced the smallest difference, and its differences were within the DTM band. As in the Dataset 1 cases, CHEQ 10 performed as well as PSE 20. As shown in Table 4, three methods (CHEQ 20/10, PSE 20) led to *RMSD*s smaller than 0.5. PEG performed nearly as well as PSE 20 in the Dataset 2A case. However,

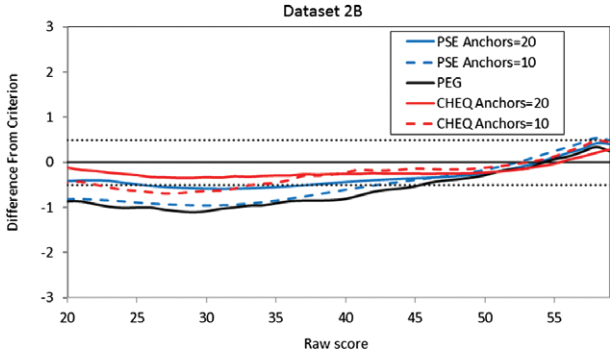

Figure 7 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 2B case.

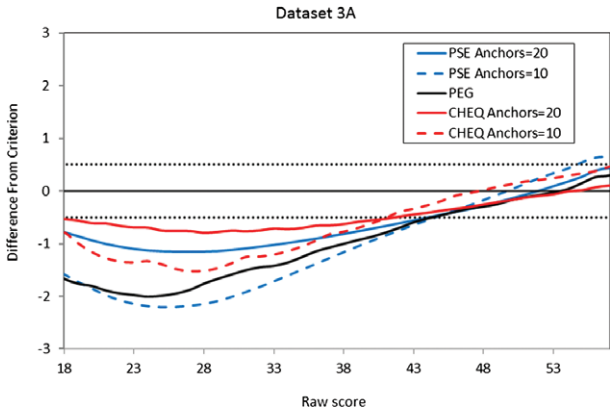

Figure 8 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 3A case.

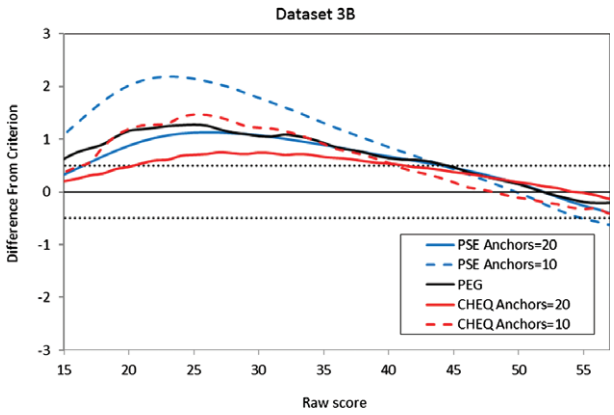

Figure 9 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 3B case.

this trend did not appear in the Dataset 2B case. Both PEG and PSE 10 produced differences larger than the DTM in the middle score region, but the magnitude of differences was generally smaller than 1, leading to *RMSD*s slightly larger than 0.5. In a similar vein, their mean deviations from the criterion mean were slightly larger than 0.5.

Figures 8 and 9 present the differences from Dataset 3A and Dataset 3B, respectively. The negative differences in Figure 8 became positive differences in Figure 9. As in the previous cases, the CHEQ 20 method produced the smallest difference. Even so, the differences associated with CHEQ 20 were beyond the DTM band across the score region where many test takers were located.[6] As a result, the overall *RMSD*s associated with all the five methods were greater

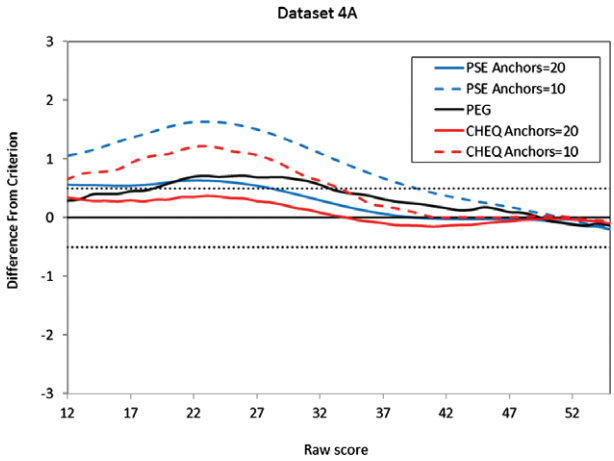

Figure 10 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 4A case.

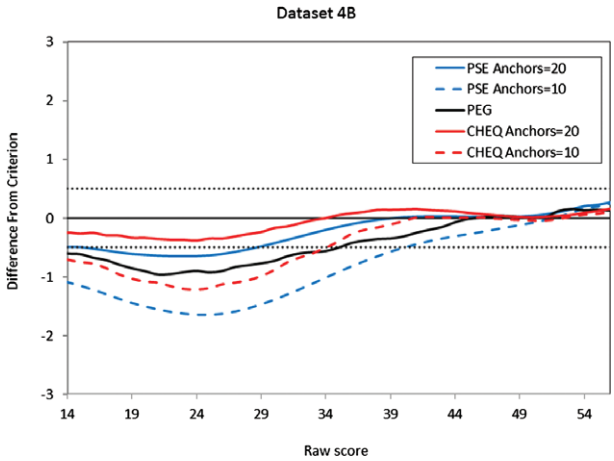

Figure 11 Equated raw score differences of poststratification equipercentile (PSE), pseudo-equivalent groups (PEG), and chained equipercentile (CHEQ) from the criterion function in the Dataset 4B case.

than 0.5 in both cases. PSE 10 performed rather similarly in both the 3A and 3B cases, but this is not true for PEG. In the Dataset 3A case, the difference pattern derived from PEG was comparable to the pattern derived from PSE 10. Both PEG and PSE 10 led to a difference greater than 1 over the middle score region (around 15 to 40). In the Dataset 3B case, however, PEG performed better than did PSE 10, leading to a smaller *RMSD* (0.754) compared to the *RMSD* (1.158) of PSE 10.

Figures 10 and 11 present the differences from Dataset 4A and Dataset 4B, respectively. The positive differences in Figure 10 became negative differences in Figure 11. In both cases, both PEG and PSE 20 performed much better than did PSE 10 and CHEQ 10, leading to the smaller *RMSD*s. The *RMSD* of PSE 10 was twice as large as the *RMSD*s of other methods. PEG performed as well as PSE 20 in both cases, and thus the differences associated with PEG were generally smaller than 1.0 across the score region from the 1st and 99th percentiles, where most test takers were located. A similar trend emerged for the equated raw score mean comparison.

Discussion

To achieve score comparability across different forms and administrations, test score equating is necessary. In some testing situations, it is difficult to implement any of the usual equating designs. In those cases, PEG linking could make it possible to report scores that are reasonably comparable across forms in a situation where solid historical information related to the target testing population is available. PEG uses test takers' background information as a major resource to achieve

Table 4 Root Mean Squared Difference of the Nonequivalent Groups with Anchor Test Equating Methods and Pseudo-Equivalent Groups Linking

Dataset	PSE (Anchor = 20)	PSE (Anchor = 10)	PEG linking	CHEQ (Anchor = 20)	CHEQ (Anchor = 10)
1A	0.349	0.696	0.744	0.220	0.432
1B	0.382	0.740	0.662	0.250	0.469
2A	0.437	0.671	0.569	0.258	0.383
2B	0.421	0.632	0.717	0.256	0.357
3A	0.689	1.097	0.948	0.505	0.720
3B	0.713	1.158	0.754	0.513	0.760
4A	0.412	1.146	0.504	0.227	0.768
4B	0.387	1.110	0.610	0.214	0.726

Note. PSE = poststratification equipercentile; PEG = pseudo-equivalent groups; CHEQ = chained equipercentile.

Table 5 Means and Standard Deviations of the Equated (or Linked) Raw Scores Derived from the Criterion and the Five Linking Methods in the New-Form Group

Dataset	Criterion Mean (*SD*)	PSE (Anchor = 20) Mean (*SD*)	PSE(Anchor = 10) Mean (*SD*)	PEG Linking Mean (*SD*)	CHEQ (Anchor = 20) Mean (*SD*)	CHEQ(Anchor = 10) Mean (*SD*)
1A	48.04 (7.69)	47.79 (7.91)	47.51 (8.10)	47.47 (8.15)	47.90 (7.84)	47.74 (7.97)
1B	46.60 (8.66)	46.88 (8.44)	47.17 (8.25)	47.14 (8.33)	46.76 (8.50)	46.93 (8.38)
2A	36.45 (10.01)	36.83 (9.83)	37.00 (9.65)	36.96 (9.81)	36.68 (9.96)	36.74 (9.78)
2B	38.54 (9.42)	38.19 (9.62)	38.05 (9.80)	37.93 (9.78)	38.31 (9.50)	38.29 (9.66)
3A	39.23 (8.65)	38.66 (9.01)	38.44 (9.37)	38.47 (9.20)	38.78 (8.86)	38.76 (9.15)
3B	37.10 (9.82)	37.70 (9.51)	37.97 (9.13)	37.74 (9.47)	37.56 (9.66)	37.64 (9.36)
4A	33.89 (10.60)	34.20 (10.35)	34.90 (10.13)	34.33 (10.42)	34.01 (10.43)	34.51 (10.20)
4B	36.14 (10.11)	35.87 (10.36)	35.18 (10.60)	35.63 (10.41)	36.06 (10.27)	35.59 (10.52)

Note. PSE = poststratification equipercentile; PEG = pseudo-equivalent groups; CHEQ = chained equipercentile.

comparability of test scores over different administrations when other alternatives for adjusting group difference in ability are infeasible or available anchor items are highly questionable. It is generally assumed that the relationship between background information and test scores is much weaker than the relationship between anchor and total test scores. This may be true for well-constructed anchor tests, but the problem is that optimal anchor tests are unavailable in some testing programs due to test security concerns. To determine the accuracy of PEG linking as an alternative for anchor equating, it is necessary to evaluate PEG linking in a situation where the true equating relationship is known. That is what this study does.

We created four pairs of research forms from the actual operational test forms and equated each pair of research forms using the five linking methods — PSE 20, PSE 10, CHEQ 20, CHEQ 10, and PEG. Because the new and reference groups differed substantially in ability by design, the CHEQ methods performed better than did the PSE methods in most conditions. CHEQ 20, which indicates an optimal (nonequivalent groups) linking condition, produced the most accurate results among the five methods across all eight conditions. This finding is not surprising. Because one third of the items were used as an anchor, the relationship between the total score and anchor score was strong. Conversely, the PSE 10 condition, which represents a weak linking condition (weak anchor and dissimilar groups in ability), led to the most inaccurate results. The PEG linking produced more desirable results than did PSE 10. PEG was an acceptable substitute for the PSE (weak anchor) equating with the four pairs of research forms used in this study. As in the simulation study conducted by Lu and Guo (2015), it appears that PEG linking could be a reasonable choice in a situation where the anchor test is questionable. When the items cannot be reused due to security concerns, practitioners may consider PEG linking as a practical option for linking scores.

Testing programs make an effort to achieve an optimal linking condition. Often, however, it is difficult to achieve an adequate condition due to a lack of common items caused by item exposure. The PEG linking design requires background variables to obtain weights that can be applied to the new-form sample to make it pseudo-equivalent to a target reference group. To enhance the effectiveness of PEG, relevant background variables, which are capable of adjusting for group differences in ability, should be collected.

In practice, background variables are mainly self-reported by test takers. Although test takers are encouraged to provide honest answers, often it is hard to ensure the accuracy of self-reported data. Perhaps some questions (e.g., gender and age) are easy to answer for most test takers. Some questions (academic major or job type) may not be always easy to answer unless the list of choices is comprehensive. Because their answers on the questionnaire are not part of their test scores, test takers are not highly motivated to answer all the questions. For that reason, many test takers simply skip the questions. Because the PEG linking depends on the quality of background data, a good strategy should be implemented to ensure the accuracy of data. For example, as a method to establish how many times test takers took the test previously, it would be better to extract this kind of information from the historical database rather than from the test taker's self- report.

Using datasets from operational administrations, this study was designed to investigate the effectiveness of PEG linking in comparison to other conventional equating methods. We assembled two research forms from each of four single forms to create situations in which the true equating relationship of the two research forms in a test-taker population is known. However, because the forms used in this study were half the length of an operational form, not an actual form, the generalization of the present finding to the full-length forms might be limited.

We manipulated differences not only in form difficulty but also in group ability to evaluate the equating approaches under situations in which NEAT equating is necessary. To define the study conditions and levels, we used our own experience in various testing programs and also the research designs from the equating literature (e.g., Kim & Livingston, 2010; Livingston & Kim, 2010). Even so, the magnitudes of differences in both form difficulty and group ability considered in the present study were substantially large and could be regarded as worst-case scenarios for test equating. It is well known that the PSE methods do not perform well with groups of widely different abilities. That PEG did better than PSE with weak anchors in a situation with substantial form difficulty and group ability differences was not surprising. Due to the limited study conditions (significant differences in both form difficulty and group ability), we were not able to determine whether PEG linking would perform as effectively as other conventional equating methods in a situation where form difficulty or group ability, or both, are fairly similar. To answer that question, a study to investigate the optimal conditions for the use of PEG in practice is recommended.

Notes

1 See Haberman (2014) for details.
2 The elements in the Z vector will be multiple independent dummy variables coded from each of the background questions on the questionnaire. In the example shown in Haberman (2015), the Z vector includes 76 independent dummy variables coded from 16 background questions.
3 The conventional use of *SMD* is to compare the group difference in ability. In this case, however, we used the *SMD* to compare the form difference in difficulty.
4 Because both groups actually took all the items in both research forms, the *SMD*s of the two groups in ability can be calculated more accurately using the total scores of the two research forms. The *SMD*s based on the total scores (60 items) were 0.18 in Dataset 1, 0.21 in Datasets 2 and 4, and 0.22 in Dataset 3. In general, the *SMD*s of the two groups tended to be underestimated when the anchor scores (20 or 10 items) were used to assess the group difference in ability. This trend was salient in the Dataset 3 case.
5 Weighting distributions based on the anchor test to create equivalent groups is essentially poststratification. Therefore, PSE and PEG are more closely related, and comparisons of them are more meaningful than the comparisons between CHEQ and PEG. Even so, it is worth noting that PSE may not be an appropriate method under the current study conditions, because by design the group ability differences are large.
6 The major reason related to the poor performance of CHEQ and PSE in the 20-common-item condition can be explained by the anchor composition. Although the *SMD* between the new and reference groups was about 0.22 (calculated using the entire 60 items), the *SMD* estimated using the 20 common items was about 0.17. For that reason, the NEAT equating methods performed poorly for Dataset 3.

References

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (Research Memorandum No. RM-94-10, pp. 93–124). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2010). Using exponential families for equating. In A. A. von Davier, *Statistical models for test equating, scaling, and linking* (pp. 125–140). New York, NY: Springer.

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information* (Research Memorandum No. RM-14-01). Princeton, NJ: Educational Testing Service.

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

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, *47*, 286–298.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practice*. New York, NY: Springer-Verlag.

Liao, C.-W., & Livingston, S. A. (2012, April). *A search for alternatives to common-item equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*, *47*, 175–185.

Lu, R., & Guo, H. (2015, April). *Comparison of PEG linking with NEAT equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Oh, H., Liu, J., & Gaj, S. (2015, April). *Application of PEG linking for testing mode adjustment in K–12 assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Xi, N., Guo, H., & Oh, H. (2015, April). *A PEG linking study of matching variables*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Suggested citation:

Kim, S. & Lu, R. (2018). *The pseudo-equivalent groups approach as an alternative to common-item equating* (Research Report No. RR-18-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12195

Action Editor: Marna Golub-Smith

Reviewers: Hongwen Guo, Rick Morgan, and Sandip Sinharay