A Simulation Study to Compare Nonequivalent Groups With Anchor Test Equating and Pseudo-Equivalent Group Linking

Ru Lu
Hongwen Guo
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.
In this paper we compare the newly developed pseudo-equivalent groups (PEG) linking method with the linking methods based on the traditional nonequivalent groups with anchor test (NEAT) design and illustrate how to use the PEG methods under imperfect equating conditions. To do this, we proposed a new method that combines the features of PEG linking and NEAT equating (referred as PEGAT) and compared it with NEAT and PEG. PEG mainly uses test takers’ background variables to create PEG and then links scores on different forms whereas NEAT equating adjusts group differences in ability through the anchor test scores. The proposed method, PEGAT, uses background variables and anchor scores to adjust group ability differences. Using simulated data, these 3 linking methods were compared in 2 equating scenarios: small and large group difference in ability. The simulation design was based on real data on a test in operation. The test scores and the background variables were assumed to have a multivariate multinomial distribution. A log linear model was used to manipulate and produce simulated data. For PEG and PEGAT linking, 3 different sets of background variables were manipulated to study the impact of correlation strength of background variables to the total scores. Our results showed that NEAT linking outperformed PEG linking when the group ability difference was large, but that NEAT linking could be improved by incorporating the PEG adjustment procedure based on background variables. When the groups were similar in ability, PEG linking produced comparable results to NEAT. This finding justifies the use of PEG linking when a good anchor test is not available as well as the use of PEGAT when a good anchor is available but needs to be strengthened by background variables.

Keywords  Linking; background variables; NEAT; PEG; simulation

doi:10.1002/ets2.12196

Equating is used to produce interchangeable and comparable scores on different test forms (Kolen & Brennan, 2004). The most popular and traditional way to perform equating is to link scores through the nonequivalent groups with anchor test (NEAT) design. In NEAT equating, the anchor test scores are used to adjust ability differences of the two samples (Kolen & Brennan, 2004). One of the fundamental requirements of NEAT design is that the anchor items need to maintain their statistical properties across the two test forms. However, this requirement might be challenging for some testing programs due to operational constraints. Though test developers aim to develop items that are fair and valid for all test takers, researchers still find the impacts of some construct irrelevant factors, such as native language and culture, on real testing situations (e.g., Abedi, 2006). Thus, when new and reference form samples perform differently on the same item, the anchor invariance assumption becomes questionable. Another major threat to NEAT equating is the test security issues. Due to worldwide access to the Internet, the anchor tests are more subject to security breaches, especially for those long anchor tests. For some high-stakes tests, same items are not allowed to appear in different forms. These psychometric or administrative restrictions make it difficult to use an anchor test.

As an alternative, PEG linking does not rely on anchor test scores. It collects test takers’ background information, such as gender, age, and educational level, as well as other variables that are related to test scores, such as learning experience and job experience. These background variables are used to create PEG by assigning weights to individual test takers in
The idea of using background variables in equating is not new. Several studies in the 1990s offered different ways of using background information with equating. For example, Lawrence and Dorans (1990) studied the effects on equating results of matching on the anchor test; Schmitt, Cook, Dorans, and Eignor (1990) studied the effects of matching on ability; Livingston, Dorans, and Wright (1990) suggested matching samples based on propensity scores; and Dorans and Wright (1993) compared equating results either matched on the equating test or matched on selection variables. The selection variable was defined as a variable or a set of variables along which subpopulation differed. Their results showed that matching on the selection variable improved accuracy over matching on the equating test for all methods. Liou, Cheng, and Li (2001) proposed to use surrogate variables (e.g., school grade, other test scores, examinee background information) with or without common-item scores to link test scores from two forms. In recent years, Brancerg and Wiberg (2011) proposed observed score linear equating with covariates, and their results showed that equating could be improved by adjusting differences in test score distributions caused by differences in the distribution of covariates. Wiberg and Branberg (2015) further applied the covariates in the framework of kernel equating and suggested that covariates could be used in practice where no anchor test scores were available.

The major difference between PEG linking and the methods mentioned above is that PEG linking utilizes the minimum discriminant information method to create two groups that can have the same background variable distributions. The number of background variables is not limited to just one or two. For example, Haberman (2015) used 76 dummy variables coded from 16 questionnaire items. None of the previous methods can manage so many background variables in the operational practice.

Several recent PEG studies have mixed results about the effectiveness of PEG linking in large scale assessments. Haberman (2015) compared the performance of PEG and NEAT using real data from 29 forms in a 2-year administration window. His results showed that PEG linking produced similar but not identical results to the NEAT operation. The overall score distributions from the PEG linking were more stable than those of the NEAT equating. Oh, Liu, and Gai (2015) compared the PEG linking with a regression-based method for mode effect in the K–12 testing context. They found that PEG linking may not be as effective as the regression-based method for mode effect. With data from a large-scale standardized test, Xi, Guo, and Oh (2015) compared PEG linking with only background variables, PEG linking with both background variables and the anchor test scores, and the operational NEAT equating. Only PEG linking with background variables and the anchor test scores could produce comparable results as the NEAT equating. They concluded that the anchor test was the single effective matching variable during the PEG weighting procedure when limited background variables are available. In these studies, PEG linking was compared with the operational procedure in which the true relationship was not known. To learn the effectiveness of PEG linking comparing to a true relationship, Lu and Guo (2015) simulated a single background variable having different degrees of casual effect with the latent ability, not the observed test scores. They found that PEG linking could outperform NEAT equating in situations where the anchor test was not sufficient. However, their simulation scenarios were too idealized. It is hard to generalize the results to operational situations. Instead of using pure simulations, Kim and Lu (2018) created two half-length research forms that shared some common items from a single operational form. Because the two research forms were from the same form, the direct equating function between the two research forms through the single group design was treated as the true relationship. Thus, PEG linking and NEAT equating were both compared with the criterion equating. They found that PEG linking produced more accurate results than NEAT design when the anchor test was short. Thus, they concluded that when anchor test is not available or not sufficient, PEG linking can be a substitute to NEAT in practice.

Except for Lu and Guo (2015), none of the other studies has evaluated the impact of the relationship between background variables and the test scores on PEG linking. An important assumption in PEG linking is that if the two groups have similar background distributions, their ability distributions would be the same. It is reasonable to expect that PEG linking with background variables that explain more variation in total test scores would produce more accurate linking results than PEG linking with background variables that explain less or little of the total scores. This study will include different background variables with PEG linking to study this effect. Instead of creating a latent variable to generate responses as in Lu and Guo, this study simulated background variables and observed scores based on a real data set.

As we mentioned earlier, NEAT linking uses anchor scores and total test scores whereas PEG linking uses the background variables and total test scores. For situations where both anchor scores and background variables are available,
neither NEAT nor PEG uses the complete available data. Therefore, a possible linking method is the combination of NEAT and PEG where all available data are used. In this study, we propose to use both the anchor test scores and the background variables to adjust group ability differences (the PEGAT linking method). We expect that NEAT equating could be improved by including the additional background information in the process of adjusting group differences in ability, especially when the total anchor correlation was low. We compare the three different linking methods (NEAT, PEG, and PEGAT) in different scenarios through simulations to evaluate whether the PEGAT method can improve the NEAT equating or PEG results and to what extent the background variables impact equating accuracy.

The purpose of this study is threefold: (a) to compare the NEAT and PEG results under different situations, (b) to investigate whether PEGAT would produce more accurate results than either NEAT or PEG, and (c) to evaluate the impact of relationship between background variables and anchor scores on PEG linking. The ultimate goal is to show how to use the PEG method to improve equating results under the imperfect equating conditions. To address the three purposes, in the “Models” section, we briefly present the NEAT design and its equating methods, the PEG linking, and the PEGAT linking. We then introduce a simulation design and evaluation methods used in subsequent sections. In this simulation design, we use loglinear models to estimate multivariate multinomial distributions based on real data in order to reflect the association of test scores and background variables observed in practice. In the “Results” section, we present comparison results of the three methods (NEAT, PEG, and PEGAT) based on simulated data from the multinomial distributions. The last section contains discussion and concluding remarks.

Models

Suppose we have two test forms, X and Y, and test takers from two samples took X and Y, respectively. Let \( x_i \) be the score of Test Taker \( i \) on Test Form X and \( y_i \) the score on Test Form Y. The test takers also take both an Anchor Test A and a questionnaire that asks about test takers’ background information such as gender, age, and education. Let \( a_i \) and \( a_j \) be the anchor test score for Test Takers \( i \) and \( j \) of Forms X and Y, respectively; let \( z_{iX} \) and \( z_{iY} \) represent the \( J \)-dimensional vector \( Z \) of background variables collected through the questionnaire for Test Takers \( i \) and \( j \) of Forms X and Y, respectively. The three linking models are illustrated as follows.

NEAT Equating Method

Multiple equating methods are used under the NEAT design (Kolen & Brennan, 2004). For this study we chose the nonlinear version of the poststratification equating (PSE) method. PSE uses the anchor test scores to estimate the distribution of \( X \) and the distribution of \( Y \) on a synthetic population, which is a mixture of the two samples on the two forms. It assumes the invariance of conditional distributions of test scores given anchor scores. Once the distributions of \( X \) and \( Y \) on the synthetic population are determined, linking methods for randomly equivalent groups are used to convert scores on the new form to scores on the old form. The reason we considered the PSE method in this study rather than other methods, such as chained equipercentile, is that PEG can be considered a poststratification method as well, which assumes the conditional invariance of test scores given background information.

PEG Linking

There are several steps in PEG linking. The first step is to obtain target background distributions to created PEG. In Haberman (2015), the weighted average background vector \( \bar{Z} \) across 29 administrations was used to represent the target background distribution. In the \( Z \) vector, 76 independent dummy variables were coded from 16 questionnaire questions. In the case of only two forms, the background means of the old form sample can be treated as the target background mean. That is,

\[
\bar{Z} = \sum_{i=1}^{N_Y} z_{iY} / N_Y
\]

where \( N_Y \) is the number of test takers on Form Y. Using the adjustment through minimum discrimination information method (MDIA; Haberman, 1984), weight \( w_{iX} \) is assigned to Test Taker \( i \) on Form X so that weighted background vector will equal to \( \bar{Z} \). That is,
where \( w_{ix} > 0 \) and \( \sum_{i=1}^{N_x} w_{ix} = 1 \). The MDIA approach uses the Newton–Raphson method to obtain the individual weight \( w_{ix} \) (Haberman, 1984).

The second step in PEG linking is to form the target test score distribution. This can be accomplished by the advanced use of exponential family (Haberman, 2010) or as in Haberman (2015), by creating the target scale score distributions by pooling the 29 operational equated scale scores together.

The third step is to conduct the equivalent group linking using the equivalent group equating method. The only difference is the weighted raw score distributions are used on Form X. With the score distributions on Forms X and Y are known, any equivalent group linking method such as mean-sigma method or equipercentile linking can be carried out. We chose the equipercentile method (Kolen & Brennan, 2004) because of the large sample size.

**The PEGAT Equating**

As in Step 1 of PEG linking, weights are obtained for the new form sample to make it pseudo-equivalent to the reference form sample. The weights are then applied to both the total test scores and the anchor scores of the new form. NEAT equating is then carried out with the newly obtained weighted new form total test and anchor scores and the reference form total and anchor test scores. To be consistent with the selected method for NEAT equating, the nonlinear version of PSE method was used in PEGAT along with PEG. This method is of particular interest when the anchor scores and the total test scores are not strongly correlated but the anchor invariance is still true, such as when the anchor test length is short.

In this study, for NEAT and PEGAT linkings, the raw scores on the total test and the anchor test were presmoothed by preserving six moments of univariate and one moment of cross-product in order to overcome data sparseness issues at some score points (Holland & Thayer, 2000).

**Method**

We simulated two equating scenarios: one when group difference in ability is small; the other when the group difference in ability is considered large. In the case of small group difference in ability, the new form sample and the reference sample were generated from the same population distributions; in the case of large group difference in ability, the new form sample and the reference sample were generated from different subpopulation distributions. To evaluate the equating results, we needed to know the true equating relationship between the new form and the reference form. In this study, for simplicity, the same form was used as both the new form and the reference form so that the true equating function was the identity line. Any equating difference from the identity line was the equating bias. We replicated the equating 100 times to compute the estimates of the equating bias and the root mean squared error of each condition.

**Population Distributions and Anchor Test Length**

To simulate data as real as possible, we referred to a foreign language test that included 100 multiple choice items and had been taken by more than 60,000 test takers. We selected a subset of items as the internal anchor to conduct NEAT equating. The long anchor test had 21 items, which meets the general minimum requirement of an anchor test; the short anchor test had only 10 items, which is usually deemed as insufficient for a test that has 100 items. The number of correct responses was used as the scores for the total and anchor tests. The total test score and two anchor test score distributions are presented in Figure 1. Along with the language test, test takers also filled out a questionnaire that produced background variables. The two background variables we selected were the experience of foreign country living and the years of the foreign language study, denoted as C and S, respectively. Variable C was scored as \( C = 0 \) (no, no experience at all) and \( C = 1 \) (yes, some foreign language living experience). Variable S was scored as \( S = 1, 2, 3, \) or \( 4 \); the larger the value \( S \) is, the more years the test taker has studied the foreign language. The relationship between the test scores and the two background variables is presented in Table 1. The mean test scores were higher for test takers with foreign country living experience.
and also higher for those with more years of foreign language study. For the total sample, the correlation between the total test score and the long anchor test score was .87; the short anchor total correlation was .79.

Multiple regression analysis was used to test if the two background variables significantly predicted test takers’ total scores. The results of the regression indicated the two predictors together explained 16% of the variance, $R^2 = .16$, $F(4, 66205) = 3111.58, p < .0001$. If regressed on each of the two variables alone, the variance explained would be 5% and 13%, respectively, for C and S. The foreign country living experience variable significantly predicted total test scores ($\beta = .23, p < .0001$), as did the years of foreign language study ($\beta_2 = .04, p < .0001; \beta_3 = .24, p < .0001; \beta_4 = .50, p < .0001$), if we used $S = 1$ as the base. Table 2 presents the ordered percentage of variance explained by having different independent variables in the regression model.

**Simulation Under the Scenario of Small Group Difference in Ability**

We assumed that the test scores (both the total test and the anchor test) and the two background variables C and S had a multinomial distribution for the population. To estimate the probabilities of the multinomial distribution for the population, the log-linear smoothing method (Holland & Thayer, 2000) was applied to the real data using the SAS code (SAS...
Table 2 The Percentage of Variance of the Total Test Score Explained by Regression Models (N = 66,210)

<table>
<thead>
<tr>
<th>Independent variable(s)</th>
<th>Variance explained (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C only</td>
<td>5</td>
</tr>
<tr>
<td>S only</td>
<td>13</td>
</tr>
<tr>
<td>C and S</td>
<td>16</td>
</tr>
<tr>
<td>Short anchor only</td>
<td>63</td>
</tr>
<tr>
<td>C and short anchor</td>
<td>64</td>
</tr>
<tr>
<td>S and short anchor</td>
<td>65</td>
</tr>
<tr>
<td>C and S and short anchor</td>
<td>65</td>
</tr>
<tr>
<td>Long anchor only</td>
<td>77</td>
</tr>
<tr>
<td>C and long anchor</td>
<td>77</td>
</tr>
<tr>
<td>S and long anchor</td>
<td>78</td>
</tr>
<tr>
<td>C and S and long anchor</td>
<td>78</td>
</tr>
</tbody>
</table>

Note. C = experience; S = years of study.

Institute, 2002) written by Moses and von Davier (2006). The same technique had been applied in Brancerg and Wiberg (2011) for their covariate linear equating study. Under the log-linear smoothing model, the logarithm of the probability for the combination of total test score $X = x_j$ (for $j = 0$ to 100), a certain anchor score $A = a_k$ (for $k = 0$ to 10 or 21), foreign country living experience $C = c$ (for $c = 0, 1$) and years of foreign language study $S = s$ (for $s = 1, 2, 3, 4$) is defined as

$$
\log (p_{jkcs}) = \alpha + \gamma^C C + \sum_{i=1}^{3} \gamma_i' S_i + \sum_{i=1}^{TX} \gamma_i' x_i' + \sum_{i=1}^{TA} \gamma_i' a_k' + \sum_{i=1}^{TXCS} \gamma_i^{CS} CS_i + \sum_{i=1}^{TXCA} \gamma_i^{CA} CA_i
+ \sum_{i=1}^{TA} \gamma_i^{CA} C a_k' + \sum_{i=1}^{3} \sum_{i=1}^{TX} \gamma_{iij} S_i x_j' + \sum_{i=1}^{3} \sum_{i=1}^{TA} \gamma_{iij} S_i a_k' + \sum_{i=1}^{IA} \sum_{i=1}^{IX} \gamma_i^{AX} A x_i' + \sum_{i=1}^{IA} \sum_{i=1}^{IX} \gamma_i^{AX} A a_k',
$$

where $\alpha$ is a normalizing constant that forces the sum of the probabilities to equal 1; the gammas (e.g., $\gamma^C$, $\gamma^S$, $\gamma^A$) are parameters to be estimated in the model. $TX$ and $TA$ are the number of moments fitted to the marginal distribution of $X$ and $A$, respectively; $IX$ and $IA$ define the cross moments fitted to the marginal distribution. We found that $TX = 6$, $TA = 6$, $IX = 1$, and $IA = 1$, that preserved the first six moments of the total test score and the anchor test score, as well as the first cross moment of the two score variables fits the data well. Using the estimated population probabilities on the cross classification of test scores and background variables $C$ and $S$, both samples of the new and reference forms were generated using the SAS procedure PROC SURVEYSELECT. The sample size for both new and reference forms was set at 60,000 to be similar to the original sample size.

Simulation Under the Scenario of Large Group Difference in Ability

The same questionnaire also asked the test takers whether they had taken the test before. History data have shown that repeaters on the test tend to score higher than first-time test takers. Therefore, to create large group difference in ability, we selected more repeaters on one sample to make it a strong (high ability) group and more first-time test takers on the other sample to make a weak (low ability) group. The same practice of manipulating group difference in ability had been applied in Kim and Lu (2018). To avoid the possible impact of unbalanced sample size on the linking results, the two groups had about the same sample size. The score means on the total test and two anchor tests of the two samples are shown in Table 3. The standardized mean difference on the overall test between the two groups (strong and weak) was about .21. For the strong group, the correlations between the total test score and the anchor test scores for the long and short anchors were .87 and .78, respectively; they were .88 and .80, respectively for the weak group.

As in the scenario of small group difference in ability, a multinomial distribution is assumed for the subpopulation distributions. We used the same procedure to estimate the probabilities of the multinomial distribution for each subpopulation. The two sets of estimated probabilities of the multinomial distributions were then used to simulate test takers’ data. The sample size of the new and old forms was set at 30,000 to be similar to the sample size in the real data.
Table 3  Test Score and Anchor Test Score Means for the Scenario of Large Group Difference in Ability

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>Years of foreign language study</th>
<th>Total test score</th>
<th>Long anchor test score</th>
<th>Short anchor test score</th>
<th>Total test score</th>
<th>Long anchor test score</th>
<th>Short anchor test score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>S = 1</td>
<td>46.53</td>
<td>10.22</td>
<td>4.69</td>
<td>52.54</td>
<td>11.42</td>
<td>5.31</td>
</tr>
<tr>
<td></td>
<td>S = 2</td>
<td>48.46</td>
<td>10.75</td>
<td>4.84</td>
<td>54.38</td>
<td>11.99</td>
<td>5.48</td>
</tr>
<tr>
<td></td>
<td>S = 3</td>
<td>53.23</td>
<td>11.70</td>
<td>5.30</td>
<td>58.26</td>
<td>12.71</td>
<td>5.87</td>
</tr>
<tr>
<td></td>
<td>S = 4</td>
<td>60.15</td>
<td>12.87</td>
<td>6.05</td>
<td>66.37</td>
<td>14.18</td>
<td>6.64</td>
</tr>
<tr>
<td>Weak</td>
<td>S = 1</td>
<td>42.81</td>
<td>9.65</td>
<td>4.22</td>
<td>48.01</td>
<td>10.77</td>
<td>4.91</td>
</tr>
<tr>
<td></td>
<td>S = 2</td>
<td>45.02</td>
<td>10.18</td>
<td>4.46</td>
<td>49.73</td>
<td>11.13</td>
<td>5.01</td>
</tr>
<tr>
<td></td>
<td>S = 3</td>
<td>50.96</td>
<td>11.29</td>
<td>5.07</td>
<td>56.67</td>
<td>12.40</td>
<td>5.70</td>
</tr>
<tr>
<td></td>
<td>S = 4</td>
<td>58.82</td>
<td>12.67</td>
<td>5.88</td>
<td>65.69</td>
<td>14.05</td>
<td>6.57</td>
</tr>
</tbody>
</table>

Background Variables in PEG and PEGAT Linkings

We tried three different adjustments with background variables: adjustment based on the variable C only, adjustment based on the variable S only, and adjustment based on both C and S. These adjustments resulted in three different PEG linkings and three different PEGAT linkings.

In all, a total of 28 simulation conditions were carried out (seven linking methods, two population designs, and two anchor test lengths). Each condition was replicated 100 times. All simulations were carried out in SAS. Particularly, PEG was run by using the SAS program noted in Haberman (2014).

Evaluation Criterion

In the simulation study, the new form and the reference form were the same. Therefore, the true equating relationship between the two forms was the identity line. With the identity line, we considered the following criteria when evaluating the outcomes of NEAT, PEG, and PEGAT.

Equating Bias

Mean equating bias: At each raw score point \( k \), the equating mean bias over the 100 replications is defined as

\[
\frac{1}{100} \sum_{j=1}^{100} \left( \hat{\tau}_{jk} - k \right),
\]

where \( \hat{\tau}_{jk} \) is the equated raw-to-raw score obtained through a certain equating method (NEAT, PEG, or PEGAT).

Weighted equating bias: For a certain equating method, the weighted equating bias over all raw score points would be

\[
\frac{1}{101} * \frac{1}{100} \sum_{j=1}^{100} \sum_{k=0}^{100} \left( \hat{\tau}_{jk} - k \right) f_j(k),
\]

where \( f_j(k) \) is the observed relative frequency at the new form raw score point \( k \) at \( j \)th replication.

Root Mean Square Error

Mean Root Mean Square Error (RMSE): For each simulation condition, the mean RMSE at each raw score point \( k \) over the 100 replications is

\[
\sqrt{\frac{1}{100} \sum_{j=1}^{100} \left( \hat{\tau}_{jk} - k \right)^2}
\]

Weighted RMSE: for each equating method, the weighted RMSE over the 100 replications for all raw score points is defined as

\[
\sqrt{\frac{1}{101} * \frac{1}{100} \sum_{j=1}^{100} \sum_{k=0}^{100} \left( \hat{\tau}_{jk} - k \right)^2 f_j(k)}.
\]
Results

Depending on the simulation conditions (small or large group difference in ability, with a long or short anchor test), each pair of data sets (new and reference forms data) was equated using seven different methods (NEAT, PEG with C, PEG with S, PEG with C and S, PEGAT with C, PEGAT with S, and PEGAT with C and S). The results are presented here for the small group difference in ability first and then for large group difference in ability.

Scenario of Small Group Difference in Ability

Figures 2 and 3 show the mean bias at each raw score point that was plotted for the long and short anchor tests for the scenario of small group difference in ability, respectively. They were obtained from Equation 2. Although we used different colors and line shapes to indicate different linking method in these two plots, almost no visible difference was present for all seven methods, except for raw scores below 20. Because few test takers scored below 20 on the total test (as shown in Figure 1), we were not concerned about the biases for raw scores below 20. For other raw score points, biases were close to the zero line, indicating the presence of negligible bias no matter which equating method was chosen. Figures 4 and 5 show the RMSE at each raw score point for the long and short anchor tests, respectively. In each plot, the lines visibly belong to two families: equating with or without anchor scores. Equatings with the anchor test (i.e., NEAT, PEGAT with C, PEGAT with S, or PEGAT with C and S) had similar magnitudes of RMSE; the three PEG linkings (i.e., PEG with C, PEG with S, or PEG with C and S) had similar magnitudes of RMSE. Equatings with the anchor test produced smaller RMSE than PEG linkings for almost all of the raw score points. However, the RMSEs of PEGAT linkings were close to that of the NEAT equating. Therefore, PEGAT did not improve NEAT equating significantly in the case of small group difference in ability. Comparing Figures 4 and 5, the equatings with the long anchor test produced smaller RMSE than the equatings with the short anchor test. This is expected as we know that a long anchor test is more reliable than a short anchor test. The PEG linkings had a similar magnitude of RMSE, indicating that PEG linkings were not impacted by the anchor test length. This finding also met our expectation because PEG linking did not utilize the anchor information at all.

Table 4 presents the weighted bias and weighted RMSE for the long and short anchor tests, respectively, for the case of small group difference in ability. It shows that the weighted bias was generally close to zero and that the weighted RMSE was small (less than .2) for all of the equating methods. Comparing to the above unweighted measures in Figures 4 and 5, other observations are similar. That is, with the same anchor test, the three PEG linkings had similar amounts of RMSE and they were generally larger than those of the NEAT equating or PEGAT linkings; the three PEGAT linkings had similar amount of RMSE as the NEAT equating. The long anchor produced smaller bias and RMSE than the short anchor for each of the equatings that utilized the anchor test scores (i.e., NEAT, PEGAT with C, PEGAT with S, or PEGAT with C and S).
The equating methods that did not use the anchor test information (i.e., PEG with C, PEG and S, and PEG with C and S) had similar amount of weighted bias and RMSE.

In summary, this simulation study indicated that, with the small group difference in ability, PEG linking produced a similar amount of bias to the NEAT equating. However, it did have more equating error than the conventional NEAT equating. PEGAT linking did not improve the NEAT equating by including the extra background variable(s) during the adjustment, nor did it deteriorate the NEAT equating.

The Scenario of Large Group Difference in Ability

Figures 6 and 7 show the mean bias at each raw score point for the scenario of large group difference in ability with the long and short anchor test, respectively. They were negative for all methods in the middle of the scale because the strong
group was used as the new form sample and the weak group was used as the reference form sample in the simulation. For the relatively large and positive new and reference group differences, it is reasonable to expect negative equating bias (e.g., Kolen & Brennan, 2004). Unlike results in Figures 2 and 3, which show that all methods had similar and close to zero bias for raw scores above 20, biases with large group difference in ability were not close to the zero line except in the top scores. PEG linkings (with C, with S, and with C and S) with the large group difference in ability were consistently larger than those of NEAT equating and PEGAT equatings, especially for scores in the middle of the scale where most test takers were. The line of NEAT equating is between PEG linkings and PEGAT linkings. Figures 6 and 7 show that PEGAT with S and PEGAT with C + S had the smallest bias along the whole scale, whereas NEAT equating was close to PEGAT linking with Variable C. The variable C added little value to the PEGAT linking or the NEAT linking. Comparing Figures 6 and 7, we found that NEAT equating and PEGAT generally had smaller magnitude of bias in the long anchor set than the biases in the short anchor set. The PEG linking was not impacted by the anchor test length. Figures 8 and 9 present the RMSEs at each raw score point for the scenario of large group difference in ability for long and short anchors, respectively. They show patterns similar to those in the bias plots, except that the RMSEs are positive.

Table 5 shows the summary weighted bias and weighted RMSE for the scenario of large group difference in ability. The results were generally consistent with Figures 6–9. For the four methods that utilized anchor test scores (NEAT, PEGAT with C, PEGAT with S, and PEGAT with C and S), long anchor tests produced smaller weighted bias and RMSE. PEG linkings were not impacted by the anchor test length. With the same anchor test length, the order of the magnitude of weighted bias of the seven linkings was the same as the order of the magnitude of weighted RMSE: From smallest to

Table 4 Weighted Bias and Weighted Root Mean Square Error (RMSE) for the Scenario of Small Group Difference in Ability

<table>
<thead>
<tr>
<th>Linking method</th>
<th>Weighted bias</th>
<th>Weighted RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Long anchor</td>
<td>Short anchor</td>
</tr>
<tr>
<td>NEAT</td>
<td>0.005</td>
<td>0.007</td>
</tr>
<tr>
<td>PEG with C</td>
<td>0.001</td>
<td>0.013</td>
</tr>
<tr>
<td>PEG with S</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>PEG with C+S</td>
<td>-0.002</td>
<td>0.007</td>
</tr>
<tr>
<td>PEGAT with C</td>
<td>0.004</td>
<td>0.007</td>
</tr>
<tr>
<td>PEGAT with S</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>PEGAT with C+S</td>
<td>0.003</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Note. NEAT = nonequivalent groups with anchor testing; PEG = pseudo-equivalent groups; C = experience; S = years of study; PEGAT = PEG and NEAT combination.
largest, they are PEGAT with C and S, PEGAT with S, PEGAT with C, NEAT, PEG with C and S, PEG with S, and PEG with C. Within the three PEGAT linkings or the three PEG linkings, the order of weighted bias or RMSE is consistent with the order of variance explained in the multiple regression models: C and S, S, and C, from the largest to the smallest as shown in Table 2.

We observed that when we only used one or two background variables in PEG linking, both weighted bias and RMSE were considerably large. With the anchor test (even if it was only 20% or 10% of the total test in this study), the weighted bias and RMSE were reduced to at least half the size of PEG linking; when all available information was used (both the background variables and the anchor test), the bias were further reduced. For example, for the long anchor test, the weighted bias of PEGAT with C and S was only about 60% of the weighted bias of the NEAT equating with the long anchor test, which was about 32% of the weighted bias of PEG with C linking. The weighted RMSE of PEGAT with C and S was about 72% of the weighted RMSE of NEAT equating, which was about 35% of the weighted RMSE of PEG with C linking. Short anchor tests showed similar patterns.
Therefore, the results show that when the NEAT design had large group difference in ability and when the anchor test functioned well and scores were available, NEAT equating was preferred over PEG linking. Here, we emphasize that the anchor test functioned as it was supposed to do. For example, the same anchor item functions in the same way for the new and reference form samples. However, occasionally in some testing programs, anchor items show differential item functioning due to group differences in culture, language, or other factors that are not measured by the test. In the absence of a good anchor test, PEG linking can be done with available background variables. In addition, a set of background variables that can better explain the ability differences might improve PEG linking performance.

The results also show that when both the anchor test and the background variables information were available, NEAT equating could be improved by incorporating the background information to adjust ability differences, as is shown in the PEGAT method, though the improvement might be limited.
Comparing the results under the two scenarios of small and large group difference in ability, our results show that linkings with small group difference in ability performed better than linkings with large group differences in ability. All methods studied with the small group difference in ability produced similar results, including the PEG linking method. However, when the group difference in ability was large, a good anchor test was important. In our study, PEG linking with limited background variables could not produce comparable results to the NEAT equating. But NEAT equating was improved by incorporating the background adjustment.

Summary and Discussion

In this study, we compared the three linking methods, NEAT, PEG, and PEGAT, using simulations under two scenarios: when the group difference in ability was small and when the group difference in ability was large. The similarity among these three methods was that they all adjusted group ability differences first and then used equipercentile equating to link scores on different forms. However, group difference was adjusted in various ways: NEAT used anchor scores, PEG used background variables, and PEGAT used both anchor scores and background variables.

The results of this study show that whenever a sound anchor test is available, linking methods that make use of anchor scores produce better results, as expected. This is because the anchor test is constructed specifically to represent the total test in terms of the content and statistical specifications. It has the strongest correlation with the total test scores. In contrast, background variables, which are not a direct measure of the abilities, can explain only the group difference to a certain degree.

When the group difference in ability is small, PEG linking can produce results similar to those of NEAT equating. One may argue that we reached this conclusion because the new and reference form samples were generated from the same population in the simulation, thus their true group difference in ability should be zero, which would make it an equivalent group (EG) design. With EG design, there is no need for the anchor test or adjustment through background variables. However, Lyren and Hambleton (2011) noted that the EG assumptions may be violated for many reasons. Our results show that PEG linking can produce results comparable to NEAT equating when the ability difference is small. This justifies the use of PEG linking in situations when there is no access to the anchor test. Thus, to correct the systematic bias caused by the minor violation of the EG assumption, either an anchor test or background variables can be used to adjust the group difference in ability. When the anchor test is not feasible, PEG can be a substitute for NEAT equating. We also observed that PEG linking generally produced slightly larger equating error than NEAT and PEGAT equatings. This finding may be partly due to the fact that the NEAT and PEGAT methods used a smoothing procedure but the PEG linkings did not. Future studies should investigate to what extent the PEG method with smoothed score distributions could reduce equating error.

When the group difference in ability is large, the anchor test in NEAT adjusts the ability difference better than the limited background variables in PEG. However, if background information is added into the ability adjustment process through PEG, such as in the PEGAT method, NEAT equating results can be improved. For testing programs that seek to improve their equating practice, this finding is particularly encouraging because sometimes the background information is not hard to get. For the testing program mentioned in this study, the background information is routinely collected during each administration (test takers complete a questionnaire after they finish the test).
We also compared PEG linking with three different sets of background variables and PEGAT with three different sets of background variables. The three sets of background variables varied in their relationship with the total test score. The results show that PEG linking is improved when the background variables explain more total test score variation. The best situation is to include all available background variables in linking, as shown in Haberman (2015), which produces more satisfactory results than using limited background variables (e.g., Dorans & Wright, 1993; Liou et al., 2001). As Stuart (2010) pointed out, it is important to include all variables known to be related to scores in the matching procedure. Generally poor performance is found in methods that use a relatively small set of predictors of convenience, such as demographics only. One advantage of the PEG method is that it has the power of including as many variables as available. This feature makes PEG distinct from other linking methods that utilize regression methods with limited background variables to link test scores (e.g., Brancerg & Wiberg, 2011; Wiberg & Branberg, 2015).

For illustration purposes, we chose limited background variables in this study. We intended to evaluate the impact of background variables on the accuracy of PEG linking with respect to their strength in predicting the total scores. The two background variables we selected for the PEG linking together accounted for about 16% of the variance in the total test score. Hence, the PEG linking did not outperform NEAT even when the anchor test was short. However, this 16% of variance often represents what is observed in real testing situations. For example, in a large-scale English language test, Wei and Qu (2014) observed that the variance of total test score that could be explained by selected background variables varied from 1.8% to 21.2%. PEG linking is expected to perform better and be closer to that of NEAT equating when more variables are included in the scenario of large group difference in ability. Future studies should manipulate the percentage of variance explained by the background variables to address the research question of when PEG can be a viable option when there is large group difference in ability.

A related question is at what point does the anchor test become insufficient. The general requirement for the anchor test is statistical and content representative of anchor test. As stated in Dorans, Moses, and Eignor (2011), many factors could affect the quality of the anchor test: A shorter anchor test tends to be less reliable and thus has low correlation with the total test. Less representable content can also result in low anchor total correlations. An anchor test that is too hard or too easy skews score distributions and thus invalidates the equating relationship. Future studies should manipulate the anchor total correlation to address the question of when the anchor test becomes insufficient.

Overall, our results confirm that the PEG linking methodology can be an alternative to traditional equating methods in situations where a good anchor test is not available and/or when the group difference in ability is large. When the group difference in ability is substantially large, none of the traditional equating methods work well (Kolen & Brennan, 2004, p. 295). However, equating results with anchor test can be improved by making use of the background variables to construct more similar groups, particularly when an anchor test is available but the correlation with the total test scores (such as very short anchors) is weak. The stronger the relationship is between the background variables and the test scores, the better the results of PEG or PEGAT will be.

One limitation of our study is that we used identical new and reference forms for simplicity. This could be seen as a first step in evaluating the effectiveness of PEG in adjusting for group difference in ability without the confounding effect of form differences. However, this study could be conducted in more realistic situations where both form differences and group differences are evaluated under PEG and PEGAT methods.

Another limitation of our study is that we compared NEAT and PEG equating results for two forms only. In practice, a large number of test forms are always used for a large testing program. This issue has considerable impact on the NEAT design’s ability to maintain scale stability (e.g., Guo, 2010; Haberman & Dorans, 2011). Compared to the NEAT design, PEG linking has the advantage of using one set of background variables available for all administrations to restrain scale drift (Haberman, 2015).

Notes

1 Because of the conceptual similarity between PEG linking and PSE NEAT equating, we chose to present PSE results for all NEAT and PEGAT methods. However, we also obtained the chained equipercentile (CE) results for NEAT and PEGAT (with C, with S, and with C and S) for each set of simulated data. They showed the same patterns as those based on PSE. The results of CE are not presented but available upon request.
References


Suggested citation:

Action Editor: Gautam Puhan

Reviewers: Longjuan Liang and Sooyeon Kim

ETS, the ETS logo, and MEASURING THE POWER OF LEARNING. are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/