Listening. Learning. Leading.®

Research Report
ETS RR–15-10

Use of Jackknifing to Evaluate Effects of Anchor Item Selection on Equating With the Nonequivalent Groups With Anchor Test (NEAT) Design

Ru Lu

Shelby Haberman

Hongwen Guo

Jinghua Liu

June 2015

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

Use of Jackknifing to Evaluate Effects of Anchor Item Selection on Equating With the Nonequivalent Groups With Anchor Test (NEAT) Design

Ru Lu,[1] Shelby Haberman,[1] Hongwen Guo,[1] & Jinghua Liu[2]

1 Educational Testing Service, Princeton, NJ
2 Secondary School Admission Test Board, Princeton, NJ

In this study, we apply jackknifing to anchor items to evaluate the impact of anchor selection on equating stability. In an ideal world, the choice of anchor items should have little impact on equating results. When this ideal does not correspond to reality, selection of anchor items can strongly influence equating results. This influence does not disappear even if large examinee samples are present. Consequently, it provides a major hazard in practical use of equating. Although the effect of anchor selection does not disappear with increasing sample size, it is reasonable to expect smaller effects with test anchors with more items. To illustrate results, two examples of real equating data were evaluated using two classical equating methods. The results show that rather large effects may be associated with sampling of anchor items.

Keywords Resampling; measurement of equating accuracy; equating stability

doi:10.1002/ets2.12056

The nonequivalent groups with anchor test (NEAT) design (e.g., Holland & Dorans, 2006; Livingston, 2004) is commonly used in equating educational tests. In the NEAT design, a new test form X is administered to a sample of examinees from population *P*, and the form is to be equated to an old test form Y administered to a sample of examinees from population *Q*. The two test forms are intended to measure the same construct, but the groups are treated as nonequivalent in the sense that the distribution of proficiency in the tested construct in population *P* is not assumed to be the same as the corresponding proficiency in population *Q*. To compare the forms, samples from both populations take the same anchor set A (see Table 1). In observed-score equating, test scores for form X and form Y are equated by examination of two bivariate sample distributions. For examinees from population *P*, the sampled variables are the test score *X* on form X and the anchor score *A* on the anchor set A. For examinees from population *Q*, the sampled variables are the test score *Y* on form Y and the anchor score *A* on the anchor set A. Equating results based on the samples from populations *P* and *Q* exhibit sampling variability typically measured by the standard error of equating (SEE; Kolen & Brennan, 2004, p. 232). This measure has been extensively studied (Liou & Cheng, 1995; von Davier, Holland, & Thayer, 2004; Zu & Yuan, 2012), and it is widely used to evaluate equating designs or methods (e.g., Kim, Walker, & McHale, 2010; Puhan, 2010; Wang, Lee, Brennan, & Kolen, 2008; Zu & Liu, 2010). A number of studies have shown that equating results also exhibit variability due to the selection of anchor items. These studies have typically involved equating by item response theory (IRT) rather than observed-score equating. For example, using data from the National Assessment of Educational Progress (NAEP), Sheehan and Mislevy (1988) found that, for the method of Stocking and Lord (1983), scale transformation parameters *A* and *B* change substantially if different subsets of the common items are employed. Fitzpatrick (2008) reported that, in a state assessment program that used IRT equating, five different anchor sets with the same length and content specifications produced different student proficiency ratings: The percent of students being classified as proficient differed by 2–3%, a variability that may have large consequences to the stakeholders of the state test. Therefore, it is quite possible that slightly different equating sets may produce different equating results.

Compared to the large number of SEE studies, only a few studies have evaluated the variability of equating or linking results due to sampling of anchor items (e.g., Haberman, Lee, & Qian, 2009; Michaelides & Haertel, 2004, 2014; Monseur & Berezner, 2007; Sheehan & Mislevy, 1988; Xu & von Davier, 2010). Among these studies, only the studies

Corresponding author: R. Lu, E-mail: rlu@ets.org

Table 1 Data Collection: Nonequivalent Groups Anchor Test (NEAT) Equating Design

Population	New Form (X)	Anchor Set (A)	Old Form (Y)
New form group (*P*)	√	√	
Old form group (*Q*)		√	√

Note. √ denotes the presence of data.

of Haberman et al. (2009) and Michaelides and Haertel (2004, 2014) have focused on equating results. Although both studies assume that anchor items are chosen at random from a large infinite pool and use IRT equating, they differ in their definition of equating accuracy and in methodology to measure equating variability. The equating error defined in Haberman et al. includes two sources of error: sampling of examinees and sampling of anchor items. They employed a double jackknife, which applies simultaneously to examinees and to anchor items. Michaelides and Haertel, however, considered only the effect of selection of anchor items. Variability due to sampling of examinees is ignored. They used both the delta method and bootstrapping to estimate equating variability. All these equating studies find effects of sampling of anchor items.

Although it is easy to find studies of observed-score equating in which SEE is computed, it appears quite difficult to find studies of effects of item sampling on observed-score equating. In this paper, jackknifing is employed to examine such effects for two observed-score equating methods: poststratification equating (PSE) and chained equating (CE). This study also considers the impact of differences between populations, equating methods, and anchor test length. Previous studies have shown that equating methods and anchor test length can affect SEE, whereas differences between populations do not have much impact on SEE (e.g., Puhan, 2010; Wang et al., 2008). It is clearly of interest to evaluate these factors in terms of sampling of anchor items.

Real data from a testing program will be used to examine variability in this study, and anchor length will be manipulated to illustrate the effect of this factor. The jackknifing method and the data will be briefly described in the following two sections. Results will be obtained for the examples under study, and conclusions will be provided in the final section. This investigation will clearly have the limitation that not all observed-score equating methods are considered, and results for other testing programs may be quite different. Nonetheless, a general methodology for examination of equating results does apply to examples other than those presented here to illustrate methodology.

Method

Jackknifing

Jackknifing is a resampling technique that may be used for estimation of variances of sample statistics (Efron, 1981; Miller, 1964, 1974; Shao & Tu, 1995; Wolter, 1985). In this resampling approach, estimates are computed for all observations and for subsets of observations. The variability of estimates for subsets is then used to estimate the variance of the estimate based on all data. To describe jackknifing, consider $n \geq 2$ independent and identically distributed random variables or random vectors X_i for i from 1 to n. Let a parameter θ have an estimate $\widehat{\theta}$ based on all the X_i. Let C be a family of subsets of the integers 1 to n with $n-k$ elements for a positive integer $k < n$, and, for A in C, let $\widehat{\theta}_A$ be the corresponding estimate of θ based on all X_i such that i is in A. In jackknifing, the estimates $\widehat{\theta}_A$ for A in C are used to estimate the asymptotic variance of $\widehat{\theta}$. In the simplest case, delete-1 jackknifing, $k = 1$ and C consists of all subsets of the integers 1 to with $n-1$ elements. Let

$$\widehat{\theta}_C = n^{-1} \sum_{A \in C} \widehat{\theta}_A. \tag{1}$$

The variance of $\widehat{\theta}$ is estimated by

$$\widehat{\sigma}_C^2 = \left[(n-1)/n\right] \sum_{A \in C} (\widehat{\theta}_A - \widehat{\theta}_C)^2. \tag{2}$$

The nonnegative square $\widehat{\sigma}_C$ of $\widehat{\sigma}_C^2$ is then the estimated standard deviation of $\widehat{\theta}$ (Miller, 1964, 1974). This form of jackknifing can be applied to sampling of anchor items, although for use with passages, it is helpful to regard the

sampling as by sets of anchor items rather than by individual anchor items. In cases in which anchor items are internal, anchor items are removed from the anchor set but not from the two tests. The parameter of interest may be a mean of equated scale scores, a conversion of a given raw score to an equated raw score, or a conversion of a given raw score to an equated scaled score. In practice, there are two main areas of concern. In real life, items or groups of items are not randomly sampled due to the existence of test specifications. The large-sample approximations on which jackknifing is based are most straightforward to verify when the parameter under study can be regarded as a twice continuously differentiable function of some vector of anchor characteristics. It is reasonable to believe that the conditions for jackknifing are relatively well satisfied, although there is also some concern because of the size of typical anchor sets. The measure $\hat{\sigma}_C$ will be used as a measure of anchor stability. Smaller values are obviously preferable to larger values.

Equating Methods

To examine effects of equating method on the effect of item sampling in anchor sets, this study considers two observed-score equating methods used for a NEAT design: PSE and CE (Holland & Dorans, 2006; Kolen & Brennan, 2004). In all approaches, a raw score on the new form is converted to an equated raw score on the old form and the raw-to-scale conversion from the old form is applied to yield a conversion of the raw score on the old form to an equated scale score on the old form. In practice, the frequency tables of raw scores on both the new forms and the old forms were presmoonthed using a log linear model.

Poststratification Equating (PSE)

In this scenario, PSE uses the anchor test A to estimate the distribution of X and the distribution of Y on a synthetic population T which is a mixture of P and Q. Once the distributions of X and Y are determined on T, linking methods for randomly equivalent groups are used to convert scores on the new form to scores on the old form.

Chained Equating (CE)

Here, CE uses the anchor A as a link. It first links X to A on P, and then links A to Y on Q. The two linking functions are then chained together to produce a conversion of X to Y.

Sampling of examinees and sampling of anchor items are different sources of equating error. To compare the magnitude of equating error due to sampling of anchor items to SEE, jackknifing on examinees was employed. For SEE, no items were jackknifed from the anchor test. The examinees with each test were randomly assigned into 51 nearly equal groups. Here 51 is an arbitrary number.[1] The 51 groups were treated as if they were individual observations, and jackknifing was performed on the groups (Miller, 1964). The further step of double jackknifing (Haberman et al., 2009) was not attempted, mainly because of the very large difference in magnitude of the standard error related to sampling of anchor items and the standard error related to sampling of examinees.

Data

Two scenarios from a large-scale testing program are considered. In each scenario, two test forms are used. Each test form has 100 right-scored multiple-choice items. To avoid identifying the testing program, the normal reporting scale is transformed to be all integers from 2 to 100. Both scenarios use the NEAT design with internal anchors. Example 1 represents a situation where the difference between the new group and the reference group was small, whereas Example 2 represents a situation where the difference was large. Example 1 has fixed anchor test length, whereas Example 2 allows us to manipulate the length of the anchor test.

Scenario 1: Small Population Difference

In the first example, data came from the same district, but at two different time points (less than a year). The anchor set contained 30 items originally selected by test developers. It is referred to as A1 in this study. A total of 39,967 examinees

Table 2 Descriptive Statistics for the Nonequivalent Groups Anchor Test (NEAT) Design With Anchor Set A1 in Scenario 1

	New Form X		Old Form Y	
Characteristic	Total	Anchor A1	Anchor A1	Total
Sample size	39,967		25,158	
Number of items	100	30	30	100
Mean	65.13	19.22	19.67	65.25
Standard deviation	15.26	4.93	4.86	15.09
Correlation	.92		.93	

took the new form and 25,158 examinees took the reference form. Table 2 shows the summary statistics on the total test and on the anchor set. In this example, the new group was less able than the reference group by 0.09 standard deviations. The correlation between the anchor set and the total test was around 0.92.

Scenario 2: Large Population Difference

The new group in Example 1 was used as the reference group in Example 2. A sample of examinees from another district was used as the new group. They took the same test form at the same time. Though it was the same form, we treated them as separate forms, X and Y, in this study. This gave us the flexibility in deciding the number of items in the anchor test. The same anchor test A1 was used in Example 2. We also picked another 30 items by approximating the content and statistical specifications. The new anchor test is referred to as A2 in the study. In this example, a total of 139,592 examinees took Form X and 39,967 examinees took Form Y. Figure 1 shows their raw score distributions on the total test and the two anchor tests. The y-axis shows the percentage of examinees at each raw score level. All three scores (total test, *A1*, and *A2*)

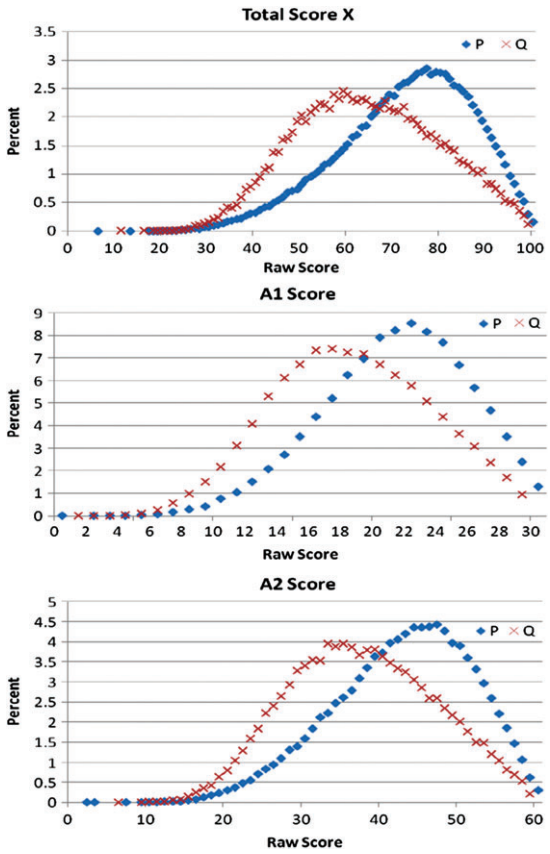

Figure 1 Raw score distributions on the total test and anchor sets in Scenario 2. P = new form sample; Q = reference form sample.

Table 3 Descriptive Statistics for the Nonequivalent Groups Anchor Test (NEAT) Design With Anchor Set A1 in Scenario 2

	New Form X		Old Form Y	
Characteristic	Total	Anchor A1	Anchor A1	Total
Sample size	139,592		39,967	
Number of items	100	30	30	100
Mean	72.52	21.02	19.22	65.13
Standard deviation	14.25	4.60	4.93	15.26
Min	6	0	2	12
Max	100	30	30	100
Correlation	.92		.92	

Note. The standard mean difference is .38 and is defined as $\left(V_P - V_Q\right) / \sqrt{\text{var}\left(V\right)_{P+Q}}$. The variance ratio is .87 and is defined as $\text{var}(V_P)/\text{var}(V_Q)$.

Table 4 Descriptive Statistics for the Nonequivalent Groups Anchor Test (NEAT) Design With Anchor Test A2 in Scenario 2

	New Form X		Old Form Y	
Characteristic	Total	Anchor A2	Anchor A2	Total
Sample size	139,592		39,967	
Number of items	100	60	60	100
Mean	72.52	42.55	38.29	65.13
Standard deviation	14.25	8.84	9.34	15.26
Min	6	2	7	12
Max	100	60	60	100
Correlation	.98		.98	

Note. The standard mean difference is .47 and is defined as $\left(V_P - V_Q\right) / \sqrt{\text{var}\left(V\right)_{P+Q}}$. The variance ratio is .90 and is defined as $\text{var}(V_P)/\text{var}(V_Q)$.

Table 5 Summary of the Three Equating Conditions in Two Scenarios

Scenario	Ability difference	Anchor	Equating Method
Scenario 1	Small	Short anchor (A1)	CE and PSE
Scenario 2	Big	Short anchor (A1)	CE and PSE
	Big	Long anchor (A2)	CE and PSE

Note. CE = chained equating; PSE = poststratification equating.

suggest that the raw scores of the new group have a distribution that is negatively skewed, whereas the scores of reference group are approximately normally distributed. Tables 3 and 4 present the descriptive statistics of examinees' raw scores on the total test and their anchors in Example 2. Both anchor sets A1 and A2 suggest that the new form group was much able than the reference group in this example. The correlations between the anchor sets and the total test were 0.92 and 0.98 for A1 and A2, respectively.

Note that in A1, the first 12 items were individual items; the other 18 items were testlet-based items. The size of a testlet was three. To avoid the possible issue of item dependence within a testlet, we used testlet as the unit for jackknifing. Every three adjacent individual items were grouped as a testlet. There were 10 testlets in A1 and 20 testlets in A2.

Thus, we have three equating conditions in the above two scenarios. They are listed in Table 5. The three equating conditions are (a) small population difference with the short anchor set A1, (b) large population difference with the short anchor set A1, and (c) large population difference with the long anchor set A2. Comparing the results of (a) and (b), we learn the impacts of population difference; comparing the results of (b) and (c), we learn the impacts of length of the anchor test. In each equating situation, two equating methods are carried out. Comparing the two equating results under the same equating condition, we can learn the impacts of equating methods. These comparisons are obviously limited. They apply to only a few forms from a particular testing program, but they do illustrate general issues.

Table 6 Total Group Mean Estimate and Its Standard Error in Scenario 1

	Anchor Set	Anchor Length	Equating Method	Average	Standard Error
Jackknife anchor item	A1	30	CE	60.39	0.56
			PSE	60.51	0.52

Note. CE = chained equating; PSE = poststratification equating.

Results

Scenario 1: Small Group Population Difference

Total Group Mean

Table 6 displays the average of the estimates of total group mean and standard errors using the jackknifing technique. The two equating methods produce similar equating results: they have similar average total group means and similar standard errors. In a typical administration, the standard error of measurement (SEM) is about 5 scale score points. Thus, the standard error of jackknifing anchor items is about 10% of SEM.

Conversions of Raw Scores to Equated Raw Scores

The detailed estimate conversions of raw to equated raw scores produced by CE and PSE using jackknifing are not presented here. The general observation is that with negative proficiency difference between populations, conversions from CE are higher than from PSE on the top and the bottom of the scale but are lower in the middle of the scale. With more examinees in the middle of the scale, it is natural that PSE has a higher average total group mean than CE. Figure 2 plots the standard error of the conversions for jackknifing anchor items. Because few examinees scored below chance level (raw score of 26), estimation of conversions is not stable, so standard errors are high. These standard errors are not plotted. In the plot, more standard errors are observed in the middle of the scale than at the two tails. The standard errors of both CE and PSE are slightly larger than 0.5 in the middle of the scale, where 0.5, half the difference between adjacent raw scores, is the difference that matters (DTM; Dorans & Feigenbaum, 1994; Holland & Dorans, 2006). This indicates that some effects of anchor item sampling are observed in the middle of the scale. Comparing the standard errors of the two equating methods, they are close to each other for most score points not at the two extremes. At all score ranges, CE is slightly higher than PSE if they differ appreciably.

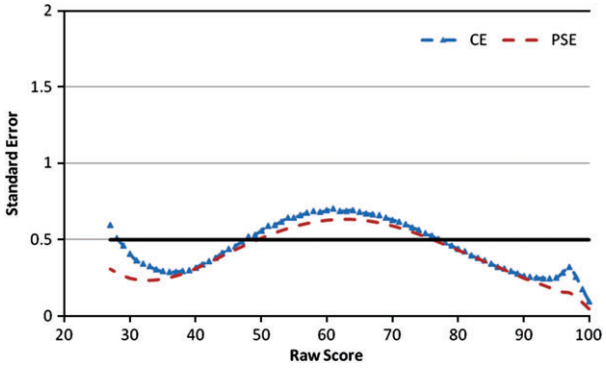

Figure 2 The standard errors of raw-to-raw conversions in Scenario 1. The blue line represents raw-to-raw conversions obtained through chained equating (CE) after jackknifing anchor items in A1 in Scenario 1; the red line represents raw-to-scale conversions obtained through poststratification equating (PSE) after jackknifing anchor items in A1 in Scenario 1.

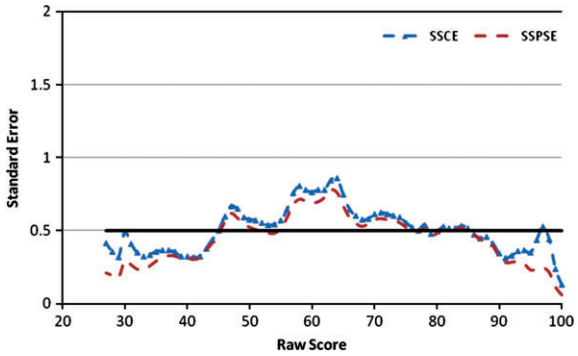

Figure 3 The standard errors of raw-to-scale conversions in Scenario 1. The blue line (SSCE) represents raw-to-scale conversions obtained through chained equating (CE) after jackknifing anchor items in A1 in Scenario 1; the red line (SSPSE) represents raw-to-scale conversions obtained through poststratification equating (PSE) after jackknifing anchor items in A1 in Scenario 1.

Conversions of Raw Scores to Equated Scale Scores

Figure 3 plots the standard errors of the conversions of raw scores to equated scale scores. Given that few examinees scored below chance level, their standard errors are not presented in the plot. Above chance level, the standard errors of raw-to-scale conversion show similar pattern as the standard errors in raw-to-raw conversions: PSE and CE produced similar standard errors of anchor item sampling at the two extremes. In the middle of the scale, CE has slightly higher standard errors than PSE. Both are larger than 0.5.

Scenario 2: Large Population Difference

Total Group Mean

Table 7 displays the average group mean estimate and its standard error using the jackknifing technique. With the same anchor test, the two equating methods produce similar total group mean estimates and standard errors. The small differences between CE and PSE with the same equating conditions are expected because of the large proficiency differences between populations. When jackknifing items, the anchor test length did have an impact on the total group mean estimate and its standard error. The total group mean estimate's difference between A1 and A2 is about 2 points, which is not trivial compared to the standard deviations of a typical administration (around 17). The long anchor test A2 has a smaller standard error for the average estimated mean of the total population. When jackknifing examinees use the complete anchor test A1, both CE and PSE give average total group means close to those from jackknifing anchor items with A1. However, their standard errors are far smaller than from jackknifing anchor items: The standard error of jackknifing examinees is only about 0.1, whereas the standard error of jackknifing anchor items is about 1. Compared to the SEM of 5, the standard error of jackknifing examinees for the estimated total group mean is about 2% of the SEM, whereas the standard error of jackknifing anchor items is about 20% of the SEM.

Table 7 Total Group Mean Estimate and Its Standard Error With Scenario 2

	Anchor set	Anchor length	Equating method	Average	Standard error
Jackknife anchor item	A1	30	CE	67.91	1.13
			PSE	67.50	1.08
	A2	60	CE	69.56	0.88
			PSE	69.40	0.86
Jackknife examinees	A1	30	CE	67.99	0.10
			PSE	67.63	0.10

Note. CE = chained equating; PSE = poststratification equating.

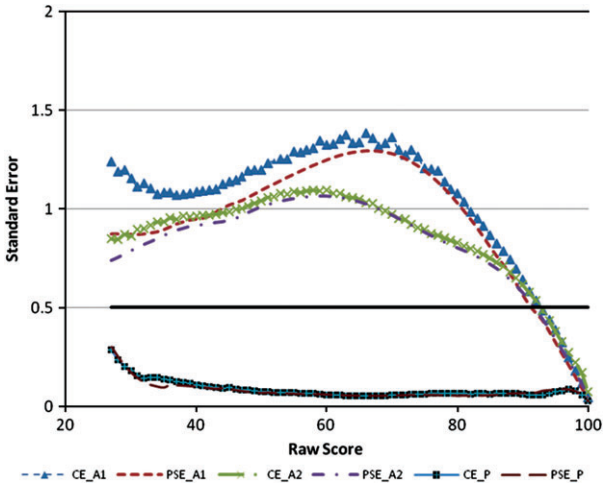

Figure 4 The standard errors of raw-to-raw conversions in Scenario 2. CE _A1 represents raw-to-raw conversions obtained through chained equating (CE) after jackknifing anchor items in A1; PSE_A1 represents raw-to-scale conversions obtained through poststratification equating (PSE) after jackknifing anchor items in A1. CE _A2 represents raw-to-raw conversions obtained through CE after jackknifing anchor items in A2; PSE_A2 represents raw-to-scale conversions obtained through PSE after jackknifing anchor items in A2. CE_P represents raw-to-raw conversions obtained through CE after jackknifing examinees using A1; PSE_P represents raw-to-scale conversions obtained through PSE after jackknifing examinees using A1.

Conversion of Raw Scores to Equated Raw Scores

Figure 4 plots the standard error of conversions of raw scores to equated raw scores for raw scores above chance level in Scenario 2. In the case of jackknifing anchor items, the standard error lines representing CE and PSE both are of parabolic shape, whereas the largest SEE is observed in the middle of the scale. This finding indicates less equating accuracy in the middle of the scale. With A1, the standard error line representing CE is constantly above the line representing PSE, indicating that the standard error of CE is slightly higher than that of PSE. With A2, the standard error lines of CE and PSE are closely together. The standard error lines (both CE and PSE) of A1 are above those of A2, indicating that A1 has a larger standard error than A2. The differences are larger in the middle of the scale. The standard error lines of jackknifing examinees (both CE and PSE) are relatively flat and close to the zero line. This indicates that the standard errors due to sampling of examinees are very small. The standard errors due to anchor item sampling of both CE and PSE under both anchor tests are larger than DTM except at the very top of the scale.

Conversions of Raw Scores to Equated Scale Scores

Figure 5 plots the standard errors of the conversions of raw scores to equated scale scores for raw scores above chance level. They show similar patterns as the standard errors in raw-to-raw conversions. Variability due to item sampling is larger with A1 than with A2. Under both A1 and A2, standard errors are large for both CE and PSE except at very high scores. When sampling examinees, the lines representing CE and PSE are flat and close to the zero line, indicating small effects of sampling examinees.

Summary and Discussion

This study applies the jackknifing methods to address the issue of anchor stability. The studied factors include the equating methods, proficiency difference, and length of anchor test. Additionally, this study compares the magnitudes of equating errors due to two sources: sampling of anchor items and sampling of examinees (Scenario 2).

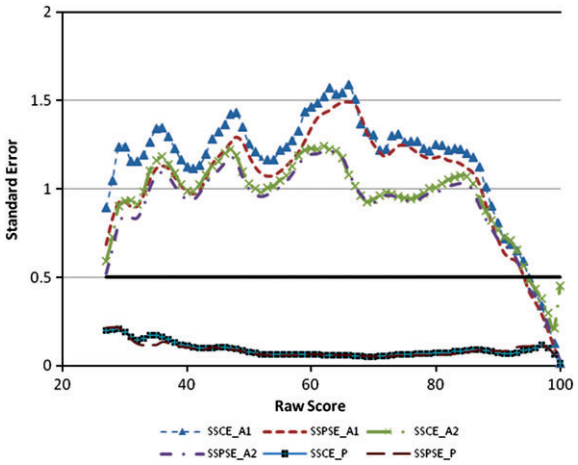

Figure 5 The standard errors of raw-to-scale conversions in Scenario 2. SSCE_A1 represents raw-to-scale conversions obtained through chained equating (CE) after jackknifing anchor items in A1; SSPSE_A1 represents raw-to-scale conversions obtained through poststratification equating (PSE) after jackknifing anchor items in A1; SSCE_A2 represents raw-to-scale conversions obtained through CE after jackknifing anchor items in A2; SSPSE_A2 represents raw-to-scale conversions obtained through PSE after jackknifing anchor items in A2; SSCE_P represents raw-to-scale conversions obtained through CE after jackknifing examinees with A1; SSPSE_P represents raw-to-scale conversions obtained through PSE after jackknifing examinees with A1.

Impact of Equating Methods

In Scenario 1, the two equating methods produce similar average estimates of total group mean and similar standard errors. In Scenario 2, with the same anchor test (A1 or A2), PSE and CE produce different average total group means but similar standard errors. For all conversions, the standard error line of CE is slightly higher than that of PSE with the same anchor test, A1. As the anchor test length increases, the standard error lines of CE and PSE become closer. Thus, we may conclude that the impact of equating methods (PSE or CE) to the standard error of total group mean estimate is limited and its impact to the standard errors of conversions is small. This finding agrees with previous findings by random sampling examinees (e.g., Wang et al., 2008).

Impact of Length of Anchor Test

In Scenario 2, the length of the anchor test was manipulated. We found that longer anchor sets produce smaller standard errors of total group mean estimates and of conversions. The difference is bigger in the middle of the scale. The results are expected because, similar to sampling of examinees, the larger the sample size is, the smaller the equating error is. A caution is that increasing the length of anchor set has its limitations in reducing the effects of sampling of anchor items. In Example 2, A2, which doubled the length of A1, still shows large effect of sampling of anchor items in the middle of the scale.

Impact of Population Differences

Scenario 1 represents a situation where the new form sample and the reference sample are from similar populations, whereas Scenario 2 represents large population differences between the samples for the new form and the reference form. Both examples come from the same testing program; however, with the same anchor set A1, the standard errors of the total group mean estimate under Scenario 2 are almost 2 times the size in Scenario 1; the standard errors of the conversions (raw to raw and raw to scale) in Scenario 2 are 3 times the size in Scenario 1 in the middle of the scale. Thus, large population differences may impact the measure of anchor stability (i.e., equating results based on large population differences may be less stable than those for small population differences). A caution is that in this study the populations may differ not only in proficiency but also in other population characteristics.

Anchor Item Sampling Versus Examinee Sampling

In Scenario 2, jackknifing was used to find the traditional SEE due to sampling of examinees. Comparison with operational equating results and with standard errors observed in Liou and Cheng (1995) indicate the reasonableness of using jackknifing to estimate the traditional SEE, although using jackknifing is more costly than the asymptotic approach. Due to the large sample sizes in both *X* and *Y* (Tables 3 and 4), the effect of examinee sampling in this study is close to zero and can be ignored when reporting scores. However, as expected, all the equating results (total group mean estimate, raw-to-raw conversion, and raw-to-scale conversion) show that in this study the standard error of anchor item sampling is much larger than that of examinee sampling. Most importantly, comparing with SEM, the effect of anchor item sampling show large effects and probably cannot be ignored in score reporting.

Two contrasts were found when comparing results of the current study to those in Michaelides and Haertel (2004, Figure 6) and Haberman et al. (2009). In our study, when jackknifing anchor items more variability was observed in the middle of the scale for the conversions for both Scenario 1 and Scenario 2. In Michaelides and Haertel (Figure 6), more accuracy was observed at the center and less accuracy was observed at the extremes. This result probably reflects use of mean/sigma scale transformation method by Michaelides and Haertel in their IRT true score equating. With the mean/sigma method, "The further a point lies from the mean of the observation, the more uncertainty there is in the prediction of the dependent variable" (Michaelides & Haertel, 2004, p. 14). Haberman et al. (2009, Table 2) apply IRT true score equating as in Stocking and Lord (1983) to much smaller samples than considered in this study, so that a substantial fraction of equating error is due to sampling examinees, and some modest additional error results from sampling of items. Overall equating errors are somewhat smaller than found in our study.

Scenario 2 illustrates the challenges of linking forms administered to very different populations. In practice, psychometricians typically conduct equating with similar proficiency and demographic distributions. The large measure of anchor stability may be a warning that large population differences are likely to be associated with large variability of equating results due to sampling of anchor. In such cases, comparison of performance of anchor items on the two forms becomes important.

Measures of anchor stability should be considered more widely than is often the case in routine equating practice. In Scenario 1, the measure of anchor stability is about 10% of SEM; in Scenario 2, the measure of anchor stability is about 20% of SEM. Although SEM is often a required element in most tests' technical reports, such a reporting practice is much less common for measurement of anchor stability. The results here show that the selection of anchor items can be an additional source of equating variability. Effects can be rather large and thus cannot be ignored in score reporting, a result consistent with Sheehan and Mislevy (1988), who found that uncertainty of the linking step can be a major source of estimation error for aggregated statistics such as total group means. Sheehan and Mislevy also used jackknifing procedures to measure the uncertainty of the linking step. However, the NAEP test in Sheehan and Mislevy and the language test in this study differ in their instruments and their population. The NAEP test is designed to measure student academic progress over time. It has a homogeneous test population from year to year. The language test is designed to measure individual student's language proficiency. As in Scenario 2, it can be taken by rather heterogeneous subpopulations. Thus, the causes for the large anchor stability effect found in Scenario 2 are not necessarily true for Sheehan and Mislevy's large variability of their linking step. Therefore, if test practitioners find large measure of anchor stability, they need to investigate possible causes and possible implications.

This study also investigated three factors that could affect the measure of anchor stability: proficiency difference, equating method, and length of anchor test. Among these three factors, equating method had the least impact in the cases studied. Length of anchor test length did have some impact. Large population difference had the largest observed impact. Factors influencing stability of anchors did not operate independently. On the other hand, the results also show that the factors are related to each other. For example, when ability difference is large, the equating method shows differences on the measure of anchor stability. The directions of the impacts are similar to those when sampling examinees.

By showing a large effect of anchor sampling in this study, especially in Scenario 2, we want practitioners to be aware of this possible source of variability in equating results. The measure of anchor stability proposed in this study can be part of a routine evaluation of the effectiveness of operational equatings. If the anchor stability is rather weak, then practitioners may need to evaluate further test assembly, test administration, and statistical analysis processes for possible causes.

Using real data, this study demonstrates the possible large effects of sampling of anchor items even when the number of examinees is quite large. Future studies can also replicate this study with equatings using small samples. Other than the three studied factors (ability difference, anchor test length, and equating method) in the study, factors such as the method of test assembly, the size of the pool of anchor items, and the contents of an item pool can also affect equating. Future studies can use simulations or other data to examine the impact of different factors (e.g., ability difference, length of anchor test, equating method, and size of item pool). More importantly, researchers and psychometricians need to find effective ways to reduce the impact of anchor item sampling.

Acknowledgments

Opinions expressed in this paper are those of the authors and not necessarily those of the Educational Testing Service or the Secondary School Admission Test Board.

Note

1 We tried different numbers and obtained similar results. Therefore, we chose 51.

References

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, I. M. Lawrence, 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.

Efron, B. (1981). Nonparametric estimate of standard error: The jackknife, the bootrap and other methods. *Biometrika*, *68*(3), 589 – 599.

Fitzpatrick, A. R. (2008). NCME 2008 presidential address: The impact of anchor test configuration on student proficiency rate. *Educational Measurement: Issues and Practice*, *27*, 34 – 40.

Haberman, S. J., Lee, Y.-H., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (Research Report No. RR-09-39). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02196.x

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187 – 220). Westport, CT: Praeger.

Kim, S., Walker, M. E., & McHale, F. (2010). Investigating the effectiveness of equating designs for constructed-response tests in large scale assessments. *Journal of Educational Measurement*, *47*, 186 – 201.

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

Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics*, *20*(3), 259 – 286.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating* (CSE Report No. 636). Los Angeles, CA: CSE and CRESST.

Michaelides, M. P., & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, *27*(1), 46 – 67.

Miller, R. G. (1964). A trustworthy jackknife. *The Annals of Mathematical Statistics*, *35*, 1594 – 1605.

Miller, R. G. (1974). The jackknife — A review. *Biometrika*, *61*(1), 1 – 15.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323 – 335.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement*, *47*, 54 – 75.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. New York, NY: Springer-Verlag.

Sheehan, K. M., & Mislevy, R. J. (1988). *Some consequences of the uncertainty in IRT linking procedures* (Research Report No. RR-88-38-ONR). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1988.tb00294.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201 – 210.

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

Wang, T., Lee, W.-C., Brennan, R., & Kolen, M. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement*, *32*, 632 – 651.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York, NY: Springer-Verlag.

Xu, X., & von Davier, M. (2010). *Linking error in trend estimation in large-scale surveys*: *A case study* (Research Report No. RR-10-10). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02217.x

Zu, J., & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement*, *47*, 395–412.

Zu, J., & Yuan, K.-H. (2012). Standard error of linear observed-score equating for the NEAT design with nonnormally distributed data. *Journal of Educational Measurement*, *49*, 190–213.

Suggested citation:

Lu, R., Haberman, S., Guo, H., & Liu, J. (2015). *Use of jackknifing to evaluate effects of anchor item selection on equating with the nonequivalent groups with anchor test* (*NEAT*) *design* (ETS Research Report No. RR-15-10). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12056

Action Editor: Matthias von Davier

Reviewers: Andreas Oranje and Jiahe Qian