Listening. Learning. Leading.®

Research Report
ETS RR–13-03

Choice of Target Population Weights in Rater Comparability Scoring and Equating

Gautam Puhan

March 2013

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Gary Ockey
Research Scientist

Donald Powers
Managing Principal Research Scientist

Frank Rijmen
Principal Research Scientist

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Director, Research

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ruth Greenwood
Editor

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.

Choice of Target Population Weights in Rater Comparability Scoring and Equating

Gautam Puhan

ETS, Princeton, New Jersey

March 2013

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

To obtain a copy of an ETS research report, please visit http://www.ets.org/research/contact.html

Action Editor: James E. Carlson

Reviewers: Mary Grant and Longjuan Liang

Copyright © 2013 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS).

Abstract

The purpose of this study was to demonstrate that the choice of sample weights when defining the target population under poststratification equating can be a critical factor in determining the accuracy of the equating results under a unique equating scenario, known as *rater comparability scoring and equating*. The nature of data collection under *rater comparability scoring* is such that it results in a very high correlation between the anchor and total score in the new form but only a moderate correlation in the reference form. I demonstrated, using data collected under a *rater comparability scoring* situation, that this difference in the anchor-total correlation in the new and reference forms can have a predictable impact on the equating results based on different sample weights (i.e., the equating results are most accurate when the reference form sample is defined as the target population, least accurate when the new form sample is defined as the target population, and somewhere in the middle when the new and reference form samples are equally weighed when defining the target population).

Key words: rater comparability scoring, target population, Tucker equating, frequency estimation equating

Acknowledgments

I would like to thank Skip Livingston for useful comments on an earlier version of the paper and Ruth Greenwood for editorial assistance.

Test equating is a statistical procedure used to produce comparable scores across parallel forms of the same test. In this paper, I focus on two poststratification equating methods, Tucker linear equating and frequency estimation equipercentile equating, which are widely used under the common item nonequivalent groups design. In Tucker equating, if new Form *X* is equated to reference Form *Y*, then the means and standard deviations of Forms *X* and *Y* are estimated for a target population and the equating is conducted by substituting these estimates into the basic formula for linear equating. In frequency estimation equating, the distributions of scores on Forms *X* and *Y* are estimated for a target population and then these estimated distributions are used to do an equipercentile equating.

A convenient way to define the target population is to make it a weighted composite of the groups taking the two forms to be equated. A target population defined in this way is called a *synthetic population* (Kolen & Brennan, 2004). To define the target population, the user can specify weights to represent the new and reference form samples in any desired proportion. For example, a weight of 1 can be applied to the new form sample and 0 to the reference form sample, in which case the target population will be the new form sample. Similarly, a weight of 0 can be applied to the new form sample and 1 to the reference form sample, in which case the target population will be the reference form sample. Another way would be to weight both the new and reference forms samples equally (i.e., $W_1 = 0.5$ and $W_2 = 0.5$).

According to Kolen and Brennan (2004), from a practical perspective, these weights have a negligible impact on the equating results. They reported results that showed that Tucker equating results were almost identical under different weighting schemes (e.g., $W_1 = 1$ versus $W_1 = 0.5$ for the new form sample). Because the choice of weights made very little practical difference, the authors recommended using a weight of 1 for the new form sample and 0 for the reference form sample.[1] According to Kolen and Brennan (1987), equating based on $W_1 = 1$ and $W_2 = 0$ allows a direct comparison of how the new sample performed on the new form to how the test takers in the new sample would have performed had they taken the reference form. Thus, such a weighting scheme led to some conceptual simplifications and was therefore desirable. Kolen and Brennan (2004) also pointed out that the equations used in Tucker equating are simplified considerably by choosing $W_1 = 1$ and $W_2 = 0$. This observation is evident in examining Equation 1 for estimating the target population variance for Reference Form *Y* (with Anchor Test

V), in which the latter part of the equation, $w_1 w_2 \gamma_2^2[\mu_1(V) - \mu_2(V)]^2$, becomes zero if $W_1 = 1$ and $W_2 = 0$:

$$\sigma_t^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1 w_2 \gamma_2^2[\mu_1(V) - \mu_2(V)]^2 . \quad (1)$$

Furthermore, if the new form sample is the target population, estimating the target population mean and variance for the new form becomes redundant, thereby reducing computational burden. Despite these advantages, Kolen and Brennan (2004) cautioned the users against using their proposed weighting scheme routinely or without any thought to any unique scenario that a testing situation might present.

The purpose of this study is to describe one such unique scenario, referred to as *rater comparability scoring and equating*,[2] and evaluate whether the choice of different weighting schemes has an impact on the equating results. Both the Tucker equating method and frequency estimation methods are evaluated**.** It seemed reasonable to focus on these two methods as they both rely on the logic of equating by conditioning on the anchor (i.e., estimating for a target population the conditional mean and standard deviation in Tucker equating or conditional distributions in frequency estimation equating and then conducting the equating using these conditional estimates).

Rater Comparability Scoring and Equating

When a multiple-choice test form is reused, the original raw-to-scale score conversion from the previous administration is typically applied to the raw scores from the current administration. However, when a constructed-response (CR) test form is reused, applying the original raw-to-scale score conversion at the current administration may not be appropriate. The difficulty of a CR test form depends not only on the items, but also on the severity of the scoring, which can vary across administrations of the same form. Rescoring provides the data to adjust for changes in the severity of the scoring. It consists of drawing a sample of responses from the previous administration and rescoring them by mixing them in with responses from the current administration. The adjustment for difficulty is based on an equating in which the new form and reference form contain the same items but are scored at different times (and, in many cases, by different raters). The new form is the test form as scored at the current administration; the reference form is the test form as scored at the previous administration. If the two sets of scores

for the same CR test form (i.e., scores obtained from the current and previous administrations) are based on scores assigned by the same number of raters (e.g., *double scoring*—two ratings of each response—on both occasions of scoring), then a single-group equating can be computed from those two sets of scores for the same papers, and the resulting conversion can be applied to the reused form in the current administration. However, due to practical constraints (cost of scoring, availability of qualified raters, etc.), testing companies often choose to double score the current operational papers but single score the papers from the previous administration (Tan, Ricker, & Puhan, 2010). In such situations, the equating results from a single-group design cannot be applied to the reused form in the current administration. Instead a *common item equating design* has to be employed, in which the anchor scores of the reused form are based on only the first rating of each response. The important principle is that the anchor scores of both equating samples must be computed by the same procedure and must be based on scores from the same scoring session (see Figure 1 for a graphical illustration of this equating model). To be consistent with equating terminology used widely, the terms *new form* and *reference form* will be used instead of *reused* and *original* forms for the remainder of the paper.

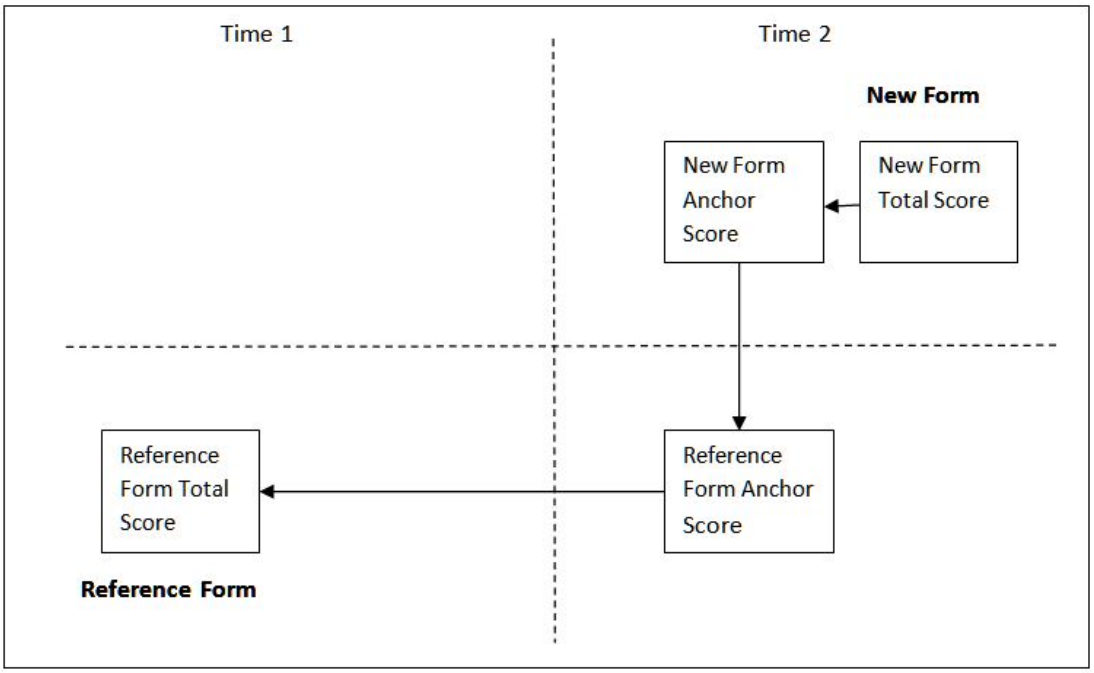

Figure 1. **Graphical illustration of the common item equating model for a new (reused) form.**

Why Is the Choice of Target Population Weights Important in This Context?

In the common item equating model described in the previous section, the relationship between the anchor and the total scores is typically much stronger in the new form than the reference form. This happens because the anchor and total scores in the new form are based on scores assigned by the same raters (i.e., by the new raters in the current administration). For the reference form, however, the anchor scores are assigned by the new raters in the current administration but the total scores are assigned by the old raters in the previous administration (i.e., the anchor score, in this case, is external to the total score). This usually results in a moderate correlation between the anchor and total scores for the reference form but a very high correlation (often close to 1.00) between the anchor and total scores for the new form.

In this context, choice of weighting schemes ($W_1 = 1$ or $W_2 = 1$) for the target population can lead to different equating results. To understand the reason for this difference, let us examine the Tucker equating formulas using a hypothetical (but exaggerated) example. The Tucker equating formulas for estimating the mean on a New Form *X* and Reference Form *Y* on a Target Population *T* are (see Kolen & Brennan, 2004, p. 108):

$$\mu_T(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)] \quad (2)$$

$$\mu_T(Y) = \mu_2(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)], \quad (3)$$

where the γ terms are the regression slopes and are defined as$\mu 2\sigma$

$$\gamma_1 = \sigma_1(X,V) / \sigma_1^2(V) \quad (4)$$

$$\gamma_2 = \sigma_2(Y,V) / \sigma_2^2(V). \quad (5)$$

In this hypothetical example where New Form *X* is equated to Reference Form *Y* using Anchor Test *V*, let $\mu_1(X) = 50$, $\mu_2(Y) = 50$, $\mu_1(V) = 20$, $\mu_2(V) = 18$, $\sigma_1(X) = 5$, $\sigma_2(Y) = 4$, $\sigma_1(V) = 2.5$, and $\sigma_2(V) = 2$. Because $\mu_1(V) > \mu_2(V)$, it is evident the new form sample is more able than the reference form sample. But the scores on the total test are the same (i.e., 50). This means that new form sample took a more difficult form (*X*) and equating should adjust for this difference in difficulty. Also, in this example, let the correlation between the anchor and total test be 0.00 in the reference form and 1.00 in the new form. If the new form sample was used as the target population ($W_1 = 1$), then the target population mean on Form *X* is simply the observed

mean on Form *X* (i.e., 50). The target population mean on Form *Y* can be estimated, using Equation 3. However, because the correlation between the anchor and total test in the reference form is 0.00, the regression term (γ_2) in Equation 3 will also become zero. Thus, the synthetic population mean for Form *Y* will simply be the observed mean for Form *Y* (i.e., 50). But because the new form is more difficult, this estimate of the synthetic population mean is not accurate and therefore an equating based on this estimate will not be accurate either. However, if the reference form sample was used as the target population ($W_2 = 1$), then the target population mean on Form *Y* is simply the observed mean on Form *X* (i.e., 50). Then using Equation 2, the target population mean on Form *X* can be estimated. Because of the perfect correlation of 1.00 between the anchor and total test in the new form sample, the anchor scores of the reference form sample provide a precise estimate of how that group (i.e., the target population) would perform on the new form. Using Equation 2, the estimated mean of the target population on Form *X* is found to be 46, which seems much more acceptable because it shows that the new form is more difficult than the reference form. The bias in Tucker equating occurs to some degree for any correlation less than 1.00; the weaker the correlation, the more the mean scores of the synthetic population are regressed to the mean of the observed samples.

The above example illustrated how the difference in the anchor-to-total correlation in the new and reference forms may affect the equating results based on different weighting schemes for the target population. With a very high correlation between the anchor and the new form, the anchor scores of the reference form sample provide a precise estimate of how that group would perform on the new form. Therefore, if the reference form sample were the target population, the equating results would be highly accurate. But with a weaker correlation between the anchor and the reference form, the anchor scores of the new form sample provide a less accurate estimate of how that group would perform on the reference form. If the new form sample were the target population, the equating results would be less accurate and biased in a predictable direction (i.e., towards the mean of the reference form sample). Finally, if the target population were an equal mix (50/50) of the new form sample and reference form sample, then the equating would not be as accurate as in the first case (i.e., $W_2 = 1$) but more accurate than in the second case (i.e., $W_1 = 1$). The purpose of this study is to evaluate this hypothesis using the Tucker linear and frequency estimation equating methods.

Method

Data and Design

The present study used data from an operational test where a criterion was available to which different equating methods could be compared. These data were collected and used in previous studies conducted by Tan et al. (2010) and Puhan (2012). The data consisted of scores assigned to examinee responses to four CR items, resulting in 48 score points (4 items $\times$ 6 maximum points per item $\times$ 2 ratings). Four hundred fifty-two (452) examinee responses from the original administration (that had been originally scored in 2005) were interspersed with 792 responses of examinees who took the same form again in 2006 and all responses were double scored by the 2006 raters. Because these 452 papers have two sets of double ratings, one set from 2005 and another set from 2006, they can be used to compute a single-group equating of the 2006 scores to the 2005 scores. For this study, the single-group equating will be considered the *criterion equating* to which the common item equatings (i.e., Tucker and frequency estimation equatings) based on different weighting schemes will be compared. Note that although all the rescored papers were double scored to facilitate the creation of the criterion equating function, only the first score will be used to mimic and evaluate the common item equating model described previously in the rater comparability scoring and equating section.

For the common item equating model, the new form total scores comprised 794 examinee responses (double scored) in the 2006 administration and the new form anchor scores comprised 794 examinee responses (single scored) in the 2006 administration. The reference form total scores comprised 452 examinee responses (double scored) in the 2005 administration and the reference form anchor scores comprised 452 examinee responses (single scored) rescored during the 2006 administration.

Equating Methods and Weighting Schemes

The common item equating methods used were the Tucker linear and frequency estimation[3] methods. There were two criterion equatings (i.e., linear and nonlinear) to which the results from the common item equating methods were compared. To evaluate the accuracy of equatings based on different weighting schemes for the Tucker equating method, the criterion equating was a single-group linear equating. To evaluate the accuracy of equatings based on different weighting schemes for the frequency estimation equating method,[4] the criterion equating was a single-group equipercentile equating. The three different weighting schemes

were: (a) a weight of 1 was applied to the new form sample ($W_1 = 1$ and $W_2 = 0$), in which case the target population was the new form sample; (b) a weight of 1 was applied to the reference form sample ($W_1 = 0$ and $W_2 = 1$), in which case the target population was the reference form sample; and (c) both the new and reference form samples were weighed equally ($W_1 = 0.5$ and $W_2 = 0.5$).

Accuracy Indices

The difference between the Tucker and frequency estimation equating functions using the different weighting schemes and their respective criterion equating functions (i.e., single-group linear or equipercentile equatings) were plotted at each score point, resulting in a difference curve for the total score region. The difference curve was also summarized into a single number known as the root mean squared difference (RMSD), which is stated as

$$RMSD = \sqrt{\sum_{j=0}^{J} w_j [e(C_j) - e(CI_j)^2]}, \quad (6)$$

where $e(C_j)$ represents the equated score at score point j using the single-group criterion equating function and $e(CI_j)$ represents the equated score at score point j derived using a particular common-item equating method (i.e., Tucker or frequency estimation equating) and weighting scheme and w_j is the weighting factor indicating the proportion of examinees at each raw score level for the criterion equating sample (i.e., $N = 452$).

Results

The means and standard deviations for the criterion and common item equating sample are provided in Tables 1 and 2. As seen in Table 1, the mean score for the 452 rater comparison papers scored in the 2005 administration is slightly higher than the mean score in the 2006 administration. The difference in the standard deviations is larger—one standard deviation being 13% larger than the other. The percentile ranks of several scores (roughly corresponding to the 10th, 25th, 50th, 75th, and 90th percentiles) in the 2005 and 2006 score distributions were quite different (see Table 3). This indicated that the scoring standards applied in 2005 and 2006 were not similar and therefore an equating adjustment was needed. As seen in Table 2, the anchor mean shows than the new form sample is less able than the reference form sample (standardized mean difference = 0.23). The correlation between the anchor and total score in the new form was

0.96 and in the reference form was 0.73. As mentioned earlier, this difference in the anchor-to-total correlation in the new and reference forms could cause the poststratification equating results following the different weighting schemes to differ.

Table 1

Summary Statistics for Single-Group Equating: 2006 Scoring to 2005 Scoring

	Examinees tested in 2005 (n = 452)	Examinees tested in 2006 (n = 452)
Score	Total	Total
Scoring in ...	2005	2006
Mean	33.53	33.44
SD	3.22	3.65

Table 2

Summary Statistics for Anchor Equating: 2006 Scoring to 2005 Scoring

	Examinees tested in 2005 (n = 452)		Examinees tested in 2006 (n = 794)	
Score	Anchor [a]	Total [a]	Total[b]	Anchor[b]
Scoring in ...	2006	2005	2006	2006
Mean	16.70	33.53	32.50	16.25
SD	1.95	3.22	3.87	2.00

[a]Correlation between the anchor and total groups in 2005 = .73. [b]Correlation between the anchor and total groups in 2006 =.96.

Table 3

Percentile Ranks of Several Scores (Corresponding to the 10th, 25th, 50th, 75th, and 90th Percentiles) in the 2006 and 2005 Score Distributions (n = 452)

Raw score	2006 (% below)	2005 (% below)
30	13.59	9.71
32	29.31	23.84
34	50.72	47.67
36	71.92	73.73
38	87.18	90.92

The difference curves comparing the linear and nonlinear poststratification equating methods using different weighting schemes are presented in Figures 2 and 3, respectively. Because equating results tend to be less stable where data are sparse, the difference was plotted for score points that fell between the 5th to 95th percentiles (i.e., raw score points 26–39) in the new form equating sample. As seen in Figure 2, the Tucker equating that defined the target population as the new form sample performed the poorest in terms of equating accuracy. The Tucker equating that defined the target population as the reference form sample performed the best in terms of equating accuracy. Finally, the Tucker equating in which the new and reference form sample were weighted equally in the target population performed better than the equating that followed the first weighting scheme but not as well as the equating that followed the second weighting scheme.

The frequency estimation equating results using the three different weighting schemes were similar to those for Tucker equating. As seen in Figure 3, the frequency estimation equating that defined the target population as the new form sample performed the poorest in terms of equating accuracy. The frequency estimation equating that defined the target population as the reference form sample performed the best in terms of equating accuracy. Finally, the frequency estimation equating in which the new form and reference form sample were weighted equally performed better than the equating that followed the first weighting scheme but not as well as the equating that followed the second weighting scheme.

The RMSDs, which summarize the information in each of the curves presented in Figures 2 and 3 into a single number, are presented in Table 4. For both the Tucker and frequency estimation methods, the equating that defined the target population as the new form sample had the largest RMSD. The equating that defined the target population as the reference form sample had the smallest RMSD. Finally, the equating that weighted the new and reference form samples equally when defining the target population had a RMSD value that was between the other two weighting schemes.

Table 4

RMSD Between the Criterion and Different Poststratification Equating Results

	Tucker	Frequency estimation
New form sample weight = 1	0.28	0.25
Reference form sample weight = 1	0.14	0.09
New and reference form samples weight = 0.5	0.21	0.17

Note. RMSD = root mean squared difference.

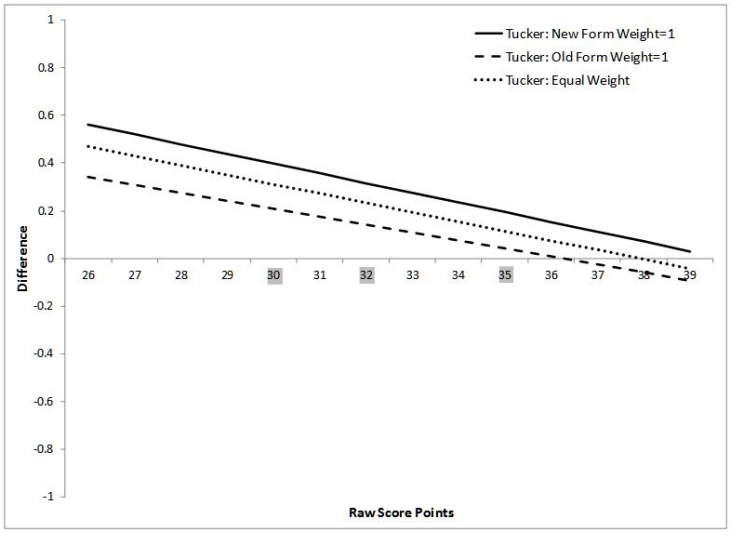

Figure 2. **Difference between Tucker equatings with three weighting schemes and the criterion. The highlighted raw score points (30, 32, and 35) indicate the 25th, 50th, and 75th percentiles.**

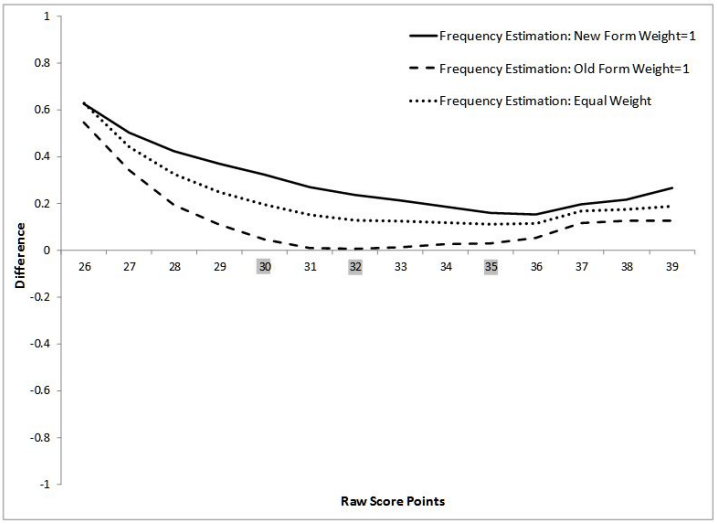

Figure 3. **Difference between frequency estimation equatings with three weighting schemes and the criterion. The highlighted raw score points (30, 32, and 35) indicate the 25th, 50th, and 75th percentiles.**

The conditional standard error of equating, which indicated variability of the equating at each score point, was also computed. The conditional standard errors of equating were fairly small for both the Tucker and frequency estimation equatings in the 26–39 raw score range. For the Tucker and frequency estimation equating methods, the largest conditional standard error of equating in this score range were 0.26 and 0.62, respectively (see Figures 4 and 5).

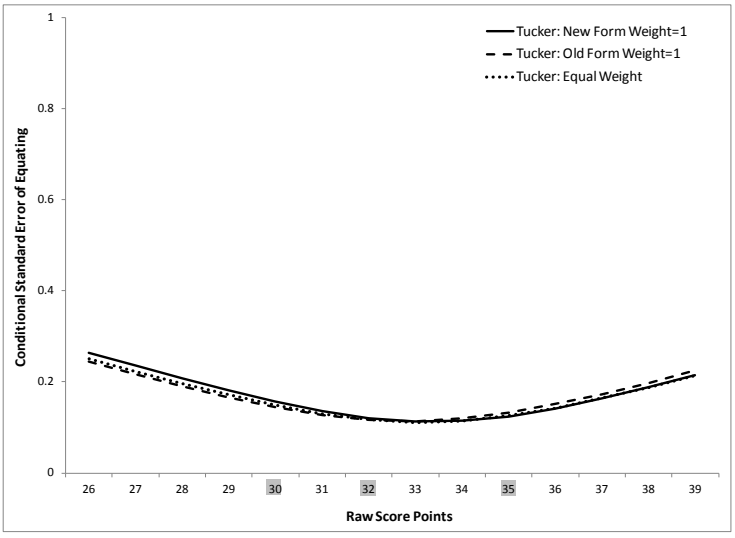

Figure 4. **Conditional standard errors for the Tucker equatings with three weighting schemes. The highlighted raw score points (30, 32, and 35) indicate the 25th, 50th, and 75th percentiles.**

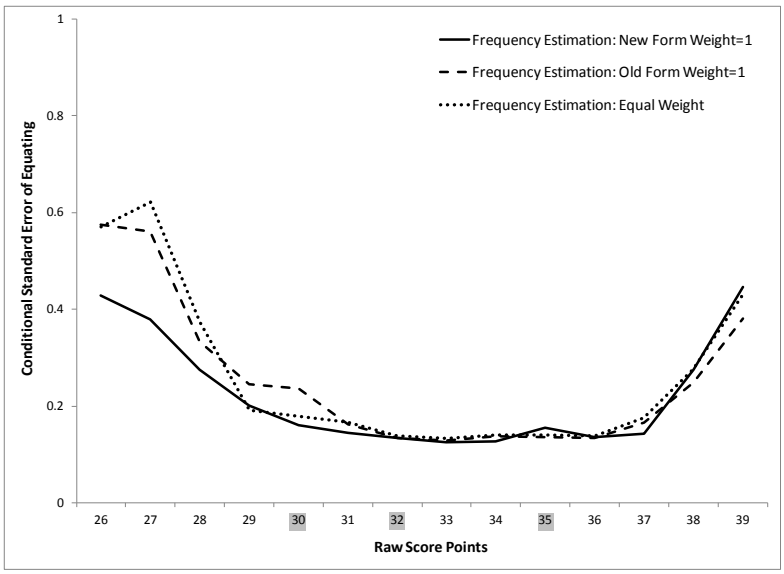

Figure 5. **Conditional standard errors for the frequency estimation equatings with three weighting schemes. The highlighted raw score points (30, 32, and 35) indicate the 25th, 50th, and 75th percentiles.**

Discussion and Conclusion

The study evaluated three different weighting schemes for defining the target population in Tucker linear and frequency estimation equipercentile equatings. The equating context in which these sample weights were evaluated is referred to as rater comparability scoring and equating, where the correlation between the anchor and total score was very high for the new form but not nearly as high for the reference form. As hypothesized, the results from both Tucker and frequency estimation equatings showed that using the reference form sample as the target population produced the most accurate results. Also as predicted, using the new form as the target population produced the least accurate results and using both the new and reference form samples in equal proportions produced an equating that was not as accurate as the equating that used only the reference form sample as the target population, but more accurate than the equating that used only the new form sample as the target population.

Based on the theoretical explanation provided earlier using the Tucker equating formulas, these results were somewhat predictable. With a very high correlation (0.96) between the anchor and the new form, the anchor scores of the reference form sample provide a precise estimate of how that group would perform on the new form. Therefore, when the reference form sample was used as the target population, the equating results were highly accurate. But with a weaker

correlation (0.73) between the anchor and the reference form, the anchor scores of the new form sample provide a less accurate estimate of how that group would perform on the reference form. Therefore, when the reference form sample was used as the target population, the equating results were less accurate. Because of the lower anchor-to-total correlation in the reference form, the estimate of how the new form sample would perform on the reference form is regressed towards the mean of the reference form sample. However, if the new and reference form samples differed in ability (as they did in this study, standardized mean difference = 0.23), this estimate would not be accurate, and an equating based on it would not be accurate. Following the same logic, the equating that uses the new and reference form samples in equal proportion when defining the target population should be less accurate than the first weighting scenario ($W_1 = 1$) but more accurate than the second weighting scenario ($W_2 = 1$).

According to Livingston (2004), when the correlation between the anchor and total scores is not very strong and there is a difference in ability between the new and reference form samples, then equating methods that rely on the logic of conditioning on the anchor (i.e., Tucker linear and frequency estimation equating) often produces biased equating results. This has also been confirmed by several empirical studies (see Puhan, 2010; Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2006). To some extent, both of these conditions were observed for the data used for this study (i.e., the correlation between the anchor and total in the reference form was moderate and the two groups differed in ability). Had the two groups been very similar in ability (e.g., such as in a randomly equivalent samples), then the differences in the equating results using the different weighting schemes would probably be negligible. Consider the previous example where the anchor-to-total correlation in the reference and new forms were 0.00 and 1.00, respectively. If the two groups were similar in ability, then in the $W_1 = 1$ condition, the anchor scores of the new form sample would not provide an accurate estimate of how that group (i.e., the target population) would perform on the reference form. In fact, because the correlation between the anchor and the total scores in the reference form is 0.00, the target population would be predicted to do as well on the reference form as the reference form examinees actually did. However, if the new and reference form samples are randomly equivalent, then this estimate would not introduce any additional bias and the resulting equating could be accurate.

The results of this study are important because they show that there are some testing situations where the choice of target population weights can impact the accuracy of equating

results. If given a choice on whether to specify the target population to be the new form or reference form sample, most practitioners would probably choose the new form sample because it is the sample for which the scores are reported. However, in a rater comparability scoring and equating situation, such a decision would result in the least accurate equating. Thus, these results serve as an important reminder that the definition of the target population for poststratification equating must depend on the particular equating situation and target population weight should not be applied routinely without a careful conceptualization of testing conditions. Finally, results from this study may also lead to a better understanding of why equating methods such as Tucker and frequency estimation equipercentile equating that use the anchor score as a conditioning variable often produces biased results when the correlation between the anchor and total test is low and when the new and reference samples differ in ability.

Implications for Practice

There are some other equating situations where the choice of target population weights may impact equating accuracy. One such situation is equating a new CR test form to a reference CR test form using rescored CR common items. Unlike rater comparability scoring and equating, the new form in this case is really a new form and not a reused form. But similar to rater comparability scoring and equating, the anchor scores for the reference form sample are computed from the rescoring and the total scores are computed from the original scoring. For the new form sample, both anchor and total scores are computed from the new scoring. Thus, the correlation between the anchor and the total score is typically low to moderate for the reference form and fairly high for the new form (see Kim, Walker, & McHale, 2010, for an example). Therefore, the current findings may generalize to the equating of new CR forms as well.

Another situation where the choice of target population weight may be important is in the event of changes made to a test to strengthen its alignment with current job requirements, curriculum changes, and so forth. If the changes are substantial, then a new scale is usually created for new test. But if the changes are small (e.g., minor changes to test specifications, formatting, etc.), then the new version of the test is typically equated to the old version of the test. In contrast to the rater comparability scoring and equating situation, in this case the anchor set (especially if it is an internal anchor) is expected to correlate better with the reference form than the new form because the anchor set would more closely match the current (reference form)

test specifications than the new test specifications. In this situation, choosing the new form sample as the target population may lead to a more accurate equating.

References

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. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*, 25–44.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement*, *11*, 263–277.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

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

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

Puhan, G. (2012). Tucker versus chained linear equating in two equating situations—Rater comparability scoring and randomly equivalent groups with an anchor. *Journal of Educational Measurement*, *49*(3), 313–330.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*, 249–275.

Tan, X., Ricker, K., & Puhan, G. (2010). *Single versus double scoring of trend responses in trend score equating with constructed response tests* (Research Report No. RR-10-12). Princeton, NJ: Educational Testing Service.

Wang, T., Lee, W., Brennan, R. L. & Kolen, M. J. (2006). *A comparison of frequency estimation and chained equipercentile methods under the common-item non equivalent groups design* (CASMA Research Report No. 17). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment.

Notes

[1] Some preliminary analyses by the author also suggested that results from the Tucker equating method using very different weighting schemes for defining the target population (i.e., $W_1 = 1$ versus $W_1 = 0$ for the new form sample) were very similar. In this case, the new and reference form samples differed in ability (standardized mean difference on the anchor was 0.26) and the correlation between the anchor and the total scores in both the new and reference forms was about 0.90.

[2] This technique has been used in the National Assessment of Educational Progress (NAEP), where it has been referred to as *trend scoring*, probably because the purpose of NAEP is to measure population trends in students' abilities. The same term was also used in research by Kim, Walker, and McHale (2010) and Tan, Ricker, and Puhan (2010). However, in this paper the term *rater comparability scoring* is used because the purpose of this scoring is to detect and control for possible changes in scoring standards of the raters, regardless of the purpose of the assessment.

[3] The reference form equipercentile equivalents of the new form integer scores obtained from the frequency estimation equating were postsmoothed using the cubic spline smoothing method described by Kolen (1984).

[4] Note that studies comparing equating results from different methods, sample size conditions, etc., often employ resampling techniques for stability of results. However, in this study the focus is more on equating bias than on random equating error. So it seemed reasonable to use one replication with the full available data.