Research Report

Impossible Scores Resulting in Zero Frequencies in the Anchor Test: Impact on Smoothing and Equating

Gautam Puhan

Alina von Davier

Shaloo Gupta

March 2008

ETS RR-08-10

Impossible Scores Resulting in Zero Frequencies in the Anchor Test: Impact on Smoothing and Equating

Gautam Puhan, Alina von Davier, and Shaloo Gupta
ETS, Princeton, NJ

March 2008

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Equating under the external anchor design is frequently conducted using scaled scores on the anchor test. However, scaled scores often lead to the unique problem of creating zero frequencies in the score distribution because there may not always be a one-to-one correspondence between raw and scaled scores. For example, raw scores of 17 and 18 may correspond to scaled scores of 150 and 153, thereby creating zero frequencies for scaled scores of 151 and 152. These gaps in the frequency distribution may adversely impact smoothing and equating. This study examines the effect of these zero frequencies on log-linear smoothing (Holland & Thayer, 1987) of score distributions and final equating results. Results suggest that although smoothing is significantly affected by the presence of these zero frequencies, as indicated by the likelihood-ratio chi-square, Akaike information criterion (Akaike, 1977), and Freeman-Tukey deviates, the impact on the actual equating results is minimal.

Key words: Test equating, log-linear smoothing, impossible scores, external anchor, scaled scores

Acknowledgments

The authors would like to thank Dan Eignor, Skip Livingston, and Rick Morgan for their helpful comments on an earlier draft of the paper and Kim Fryer for editorial assistance.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Test equating is a statistical procedure used to adjust difficulty differences across parallel forms of a test, which in turn allows for score comparisons across different groups of examinees, regardless of the test forms they were administered (Kolen & Brennan, 2004; Livingston, 2004). Equating designs often involves a single population (i.e., single-group design) or two populations (i.e., the common-item nonequivalent groups design). In the second equating design, scores on two forms of a test are linked using an anchor test (Kolen & Brennan; von Davier, Holland, & Thayer, 2004).

Equating under the common-item nonequivalent groups design is carried out using a set of items (a *mini-test*) that appear on both the new form (in population *P*) and a reference form. (in population *Q*). This type of anchor test is known as an internal anchor, where scores on the anchor test are included in the total score, and is commonly used in the case of multiple-choice (MC) tests and with some constructed-response (CR) tests. However, constructed-response tests often have fewer test items, which makes it difficult to use the common item design. For example, it is difficult to get an adequate internal anchor on a CR test which has six items, because having too many of the six items repeated raises a security concern, while having too little repeated may result in a weak anchor. In such cases, an external test (e.g., an MC test) that measures, to the extent possible, the same construct as the CR test and has been taken by the same test takers who took the CR test can be used as an external anchor to equate the new CR test form. The equating is conducted using the sub-sample (or equating sample) of CR test takers who also took the MC anchor test. Ideally, the external anchor should measure the same knowledge and skills as the CR test to be equated. In reality, it is often difficult to achieve this ideal, therefore the adjustment in form difficulty can be viewed more as a type of *linking* than as an equating process.

External Anchor and the Zero Frequency Issue

Equating under the external anchor design can be done using either raw or scaled scores on the anchor test. However, with tests having small sample sizes, scaled scores (which are comparable across different test forms) may be preferable. This is because scaled scores would allow for the accumulation of data from samples taking different forms along with the MC anchor test at previous reference form administrations Moreover, if the MC anchor used in the new form administration is also a new test form, then the only way to get anchor scores on a reference form is to use scaled scores from a different anchor form. Consider a new CR form administered in June

and equated back to an old CR form administered in January. Unless the same MC external anchor form was administered in January and June, raw scores on the MC anchor cannot be used to equate the CR test forms. However, since scaled scores are intended to be comparable across different test forms, it is possible to use scaled scores on two different MC test forms as anchor scores to equate the new and reference CR forms.

Scaled scores can often produce a unique problem of creating impossible scores (resulting in zero frequencies for these impossible scores) in the frequency distribution, because there may not always be a one to one correspondence between raw and scaled scores. Consider an anchor test with 80 items (e.g., Form X) with an established raw-to-scale conversion table. Also, consider that this conversion table is based on a linear equating of Form X to an old MC form (e.g., Form Y) and then linked to the base scale (e.g., 100 to 300 score points with increments of 1). Suppose the slope and the intercept derived from the linear equating mentioned above are 3 and 30, respectively. Using these parameters, a raw score of 20, for example, will correspond to a scaled score of 90, but a raw score of 21 will correspond to a scaled score of 93. Therefore, when scaled scores derived using this conversion table are used as anchor scores in the external anchor equating, there will be zero frequencies for scaled score points such as 91 and 92, even though the neighboring scores of 90 and 93 may have high frequencies. [1]

The purpose of smoothing in equating is to estimate the distribution that would occur in a much larger group of examinees. In the context described above, that distribution would have zero frequencies for the impossible scores. But, one can save time and effort by using a smoothing process that treats these impossible scores as if they were possible. Although it is reasonable to expect that the smoothing will not be accurate if the smoothing process treated the impossible scores as if they were possible, it could still produce accurate equating results.

Therefore, the purpose of the current study was to assess if the smoothing method that treated the impossible scores as possible will produce equating results similar to those produced by the smoothing method that treated the impossible scores as impossible. Since no previous study (to our knowledge) has investigated this issue, and since the results of the tests under investigation are used for high stakes pass and fail decisions, it was especially important to investigate this problem.

Method

Test and Data Description

The study used test data collected during the 2004-2005 administrations of two different tests using cut scores. Actual cut scores and names of the tests and score users are not identified in this paper, for confidentiality reasons. Hypothetical names for the users such as User 1 or User 2 are used. These two tests, referred to as Test A and Test B throughout the paper, assess basic skills in Elementary education and Spanish, respectively. The scaled score distributions for the anchor tests for both Tests A and B have a fairly high incidence of the impossible scores and were therefore considered adequate for this study.

Test A consisted of four CR items, resulting in 48 raw score points (4 items X 2 raters X 6 possible scores for each item). The new and old forms of Test A will be referred to as Form M and Form N, respectively. The MC external anchor test used to equate Form M and Form N will be referred to as AV. Test AV consisted of 110 items with a score scale extending over 100 points with increments of 1 and measured similar skills as Test A. Two different forms of Test AV (Forms AV_M and AV_N) were given along with Form M and Form N, respectively therefore scaled scores on these two anchor forms were used to equate Forms M and N.

Test B consisted of nine CR items, resulting in 120 raw score points (9 items X 2 raters X 4 possible scores for each item). The new and old forms of Test B will be referred to as Form Q and Form R, respectively. The MC external anchor test used to equate Form Q and Form R will be referred to as BV. Test BV consisted of 120 items, with a score scale extending over 100 points with increments of 1 and measured similar skills as Test B. Two different forms of Test BV (Forms BV_Q and BV_R) were given, along with Form Q and Form R, respectively, and therefore scaled scores on these two anchor forms were used to equate Forms Q and R.

For Test A, a large proportion of test takers who took Forms M and N also took Forms AV_M and AV_N. The new form sample that took both Form M and Form AV_M (the equating sample) consisted of 1404 examinees and the old form sample that took both Form N and Form AV_N consisted of 1735 examinees. Since the equating sample includes those test takers who took the CR test and also took the MC external anchor test, it is often smaller than the sample who took only the CR test. Similarly, for Test B, the new form sample that took both Form Q and Form BV_Q consisted of 223 examinees and the old form sample that took both Form R and Form BV_R consisted of 433 examinees.

Analytical Procedure

Equating Method

New forms were equated under the common-item nonequivalent groups design using chained equipercentile (CE) and frequency estimation equipercentile (FE) equating methods (Livingston, 2004; Kolen & Brennan, 2004). Before the actual equating relationship was derived between the new and old forms, log-linear smoothing (Holland & Thayer, 1987 was applied to the score distributions to adjust for the irregularities of the score distributions, which may cause problems for the CE or FE equating methods.

The presmoothing log-linear model used to fit the bivariate distribution of scores of any two tests (call them *X* and *Y*) is given as

$$\log(p_{jl}) = \alpha + \sum_{i=1}^{I} \beta_{Xi} (x_j)^i + \sum_{k=1}^{H} \beta_{Yk} (y_l)^k + \sum_{g=1}^{G} \sum_{f=1}^{F} \beta_{gf} (x_j)^g (y_l)^f \quad , \qquad (1)$$

where p_{jl} is the joint score probability of the score (x_j, y_l ; score x_j on test *X* and score y_l on test *Y*, for this population), the β 's are parameters that are estimated in the model-fitting process, and α is the normalizing constant selected to make the sum of p_{jl} equal to one. Note that the *X* and *Y* terms in the equation above refer to any two tests. Therefore, they can be replaced by terms such as A or AV to denote the total and anchor tests in the current study. The fitting of (1) produces a smoothed bivariate distribution that preserves *I* moments in the marginal (univariate) distribution of *X*, *H* moments in the marginal (univariate) distribution of *Y*, and a number of cross-moments ($G \leq I,\ F \leq H$) in the bivariate *XY* distribution. The moments of the log-linear models are abbreviated as I-H-G-F(M-N), where the meaning of I, H, G, F are as just explained. We will use this abbreviation in this study hereafter M and N refer to the moments of the score points excluded from the fitting of the log-linear model if a selector vector is used (e.g., when zero frequencies for impossible scores are excluded in the smoothing). We will explain this selector vector technique later in this section. See Holland & Thayer (1987) for a detailed description of this model; including algorithms for estimation, properties of the estimates, and applications to fitting test score distribution.

Equating was conducted under two smoothing conditions (referred to as Conditions 1 and 2). In Condition 1, the equating used the smoothed frequency distributions where the impossible scores were included in the smoothing. In Condition 2, the equating used the smoothed frequency distributions where the impossible scores were excluded in the smoothing (i.e., using

selector vectors to exclude the score points with zero-frequencies from the fitting of the log-linear model). The difference in the smoothing in Conditions 1 and 2 applied only to the anchor score distributions and not the total score distributions, because only the anchor score distributions based on scaled scores have the impossible zeros. The total score distributions were smoothed using the usual smoothing procedure (i.e., since zero frequencies for the impossible scores did not exist for the total score, a change in the smoothing to exclude zeros was not necessary).

A decision was made to exclude zeros (Condition 2) that were observed in the interval from plus to minus one-and-a-half standard deviation around the mean, where most of the data are. There were two reasons for doing this. First, some of the observed zero frequencies in a frequency distribution are due to sampling (e.g., very high or low scores often have real zero frequencies) and therefore, removing all zero frequencies may not be appropriate. Second, when two or more different anchor tests forms are combined (e.g., AV_{N1} + AV_{N2}) to accumulate data from different samples (especially true for the old form), the impossible zeros cannot be readily identified and removed because an impossible zero on one anchor test form may have a possible score on another anchor test form, and vice versa. Using the one and a half SD around the mean approach seemed like a reasonable solution to the problem. If accumulating data for either the new or old form samples is not necessary, only one anchor form will be associated with the new and old CR forms. In such cases, using the one and a half SD around the mean approach described above is not necessary. One can simply look up the raw-to-scale conversion table and identify which anchor scores are possible and which are not.

Log-linear modeling with selector vectors. Indicator functions can be used to fit both the full univariate distribution and a subset of the distribution (e.g., *teeth,* or lumps at different score points) within a single loglinear model. One example of such a model is:

$$\log_e(p_j) = \alpha + \beta_1(x_j)^1 + \beta_2(x_j)^2 + \beta_3 S_j + \beta_4(x_j)^1 S_j \quad , \qquad (2)$$

where the indicator function S_j =1 if *j* belongs to a defined subset of all *j's* and S_j =0 otherwise. S_j denotes the set of score points where the frequencies are systematically lower or higher than most of the test frequencies. Equation (2) will preserve the mean and variance of the total distribution of *X* (β_1 and β_2), the total frequency in the cells denoted as S_j =1 (β_3), and the mean of the cell values for the cells in S_j =1 (β_4).

One additional smoothing model combines the model in (1) with the use of indicator function in (2) to address the zero-frequency score points in the anchor (*X* and *V* are used to designate the total and anchor tests, respectively, in the equation below):

$$\log_e(p_{jk}) = \alpha + \beta_l S_l + \sum_{i=1}^{I} \beta_{xi}(x_j)^i + \sum_{k=1}^{H} \beta_{vk}(v_k)^k + \sum_{g=1}^{G} \sum_{f=1}^{F} \beta_{gf}(x_j)^g (v_k)^f + \sum_{k=1}^{H} \beta_{vkS}(v_k)^k S_l \quad . \qquad (3)$$

Equation (3) preserves the subgroups' (S_l) frequencies, *X* and *A* univariate moments, and *XA* cross-moments. The S_l term denotes that the above model allows subgroups distributions (if specified) to differ in all of their moments on *X*, all of their moments on *A*, and all of the cross-moments on *XA*.

The log-linear model with selector vectors we fitted in this study is of the form "5, 5, 1, 1" (for Test A) which, as explained above, refers to the 5 moments of *X* and *A* marginal scores, one moment of *X* and *A* in the cross products, and one moment of *A* for the selector vectors. For Test B, the log-linear model with selector vectors we fitted is of the form "3, 3, 1, 1," which refers to the 3 moments of *X* and *A* marginal scores, one moment of *X* and *A* in the cross products, and one moment of *A* for the selector vectors.

Methods used to evaluate smoothing in Conditions 1 and 2. The fit of the log-linear smoothing function to the observed data was evaluated using three numerical indexes: Likelihood ratio chi-square, the Akaike information criterion or AIC (Akaike, 1977), and Freeman-Tukey (FT) deviates. The likelihood chi-square is given as:

$$G^2 = 2\sum_j n_j \log_e\left(\frac{n_j}{\tilde{p}_j}\right), \qquad (4)$$

where $\tilde{p}_j$ is the smoothed values of p_j based on a particular model, and p_j is the population score probability of the scores in the j_{th} cell. A smaller chi-square among competing models indicates a better fit (Holland & Thayer, 2000 The AIC is given as:

$$AIC = X^2 + 2(k+1) \quad , \tag{5}$$

where X^2 is the likelihood ratio chi-square and k is the number of moments (e.g., mean, standard deviation, skewness) preserved in the smoothing. Livingston (1996) recommended using the AIC as a criterion to evaluate the efficiency of smoothing for competing models (e.g., smoothing the same data while preserving different number of moments, or smoothing with the impossible scores included versus smoothing with the impossible scores excluded). The best smoothing, as evaluated using the AIC, is the one that minimizes the value of the AIC. Finally, the formula for the FT deviates is given as:

$$\sqrt{(o(i))} + \sqrt{(o(i)+1)} - \sqrt{4M(i)+1)} \quad , \tag{6}$$

where $o(i)$ is the observed frequency in the i^{th} cell and $M(i)$ is the fitted frequency. Since the FT deviates perform roughly like independent standard normal deviates, an FT deviate with an absolute value greater than 3 would be considered large (see von Davier et al., 2004). Furthermore, when comparing the smoothing between Conditions 1 and 2, if the FT deviates in one condition across most score points tend to be smaller than the corresponding FT deviates in the other condition, it would indicate a better model fit for the condition with the relatively smaller FT deviates.

Method used to detect a difference between two equating functions. The equating results from Conditions 1 and 2 were compared both graphically and numerically. Under the graphical approach, the equating functions derived from Condition 1 were compared to Condition 2 using a difference curve, which essentially means calculating the difference between equated scores derived from the two equating conditions at each score level and plotting this difference on a graph. Although large differences are often found in the tails of the equating functions because of sparse data, particular attention was given to any difference that was found in the cut score range (the tests used in this study have cut scores with different cut scores specified by different users).

Under the numerical approach, equating functions were compared by calculating the root expected squared difference (RESD) between two equating functions. The RESD is similar to the more conventionally used root mean square difference (RMSD). The only difference between them is that the RESD weights the difference between equated scores at each score level based on the frequency of examinees at each score level. The formula for RESD is

$$RESD = \sqrt{\sum_{x=0}^{X} w_x \{[e_1(x) - e_2(x)]^2\}}, \quad (7)$$

where x represents each raw score point, e_1 represents the equated scores in Condition 1, e_2 represents the equated scores in Condition 2, and w_x is the weighting factor indicating the proportion of examinees (in population *P*) at each raw score level. Smaller values of RESD indicate smaller differences between two different equating functions.

Criteria used to evaluate a large difference between two equating functions. Dorans and Feigenbaum (1994) proposed the notion of *differences that matter* (DTM) which proposes that any difference that makes a difference in the reported score is a DTM. This is usually greater than half of the reported score unit. The DTM notion is useful for tests using cut scores where one half of the reported score unit can potentially change the pass or fail status of examinees. Since for both tests used in this study scaled scores are reported in 1 point increments, half of this unit can make a difference in pass/fail status of examinees. Therefore, a DTM of 0.5 was used for these tests to evaluate differences between two equating functions.

The *conditional standard errors of equating* or CSEE (equating error at each score point) were also used to evaluate the precision of equating in Conditions 1 and 2 (i.e., a lower CSEE would indicate a more precise equating).

Pass/Fail Decisions Based on Conditions 1 and 2 Equatings

Finally, the pass/fail status of examinees based on the different equating functions derived in Conditions 1 and 2 will be examined. This is particularly important for tests using cut scores, because even small differences in the equating functions (which may be undetected using the DTM criteria) may have a considerable impact on the pass and fail status of examinees. For example, a difference between scores such as 40.4 and 40.6 would be undetected using the DTM criteria, but 40.4 would round down to 40 and 40.6 would round up to 41. Consequently, this may lead to a difference in pass/fail status of examinees who received such scores. On the contrary, often differences larger than the DTM may exist at different score levels but they may still not affect the pass/fail status of examinees. For example, a difference between scores such as 40 and 41 is greater than the DTM of 0.5. But if the cut score is 40 then it would not matter whether an examinee received a score of 40 or 41. Both examinees would pass.

Results

All results presented below were obtained using the CE and FE methods under two different smoothing conditions (smoothing with impossible scores included versus smoothing with impossible scores excluded). Smoothing in both conditions was evaluated using the likelihood ratio chi-square, AIC, and FT deviates. Differences in equating functions derived under the two conditions were evaluated using the difference curve, RESD, DTM, CSEE, and pass and fail results.

Smoothing Results for Conditions 1 and 2

The observed score distributions in Conditions 1 and 2 were smoothed using the bivariate polynomial log-linear model. Table 1 presents the fit statistics for the smoothed score distributions for both the new and old form samples for Tests A and B. Table 2 presents the summary statistics for Tests A and B. Test A had a fairly large sample and the total and anchor score distributions were smoothed, preserving five moments. Test B had a relatively smaller sample size and fewer moments (i.e., 3 moments) were preserved in the smoothing. Note that the smoothed univariate distributions for the total and anchor scores are obtained directly from the smoothed bivariate distributions.

As seen in Table 1, the likelihood ratio chi-square for Test A was much smaller in Condition 2 than Condition 1, suggesting that the model fit was better when the impossible scores were removed from the smoothing. Similarly, the likelihood ratio chi-square for Test B was smaller in Condition 2 than Condition 1, suggesting that the model fit was better when the impossible scores were removed from the smoothing.

The AIC for Test A was smaller in Condition 2 than Condition 1, suggesting that the model fit was better when the impossible scores were removed from the smoothing (see Table 1). Similarly, the AIC for Test B was smaller in Condition 2 than Condition 1, suggesting that the model fit was better when the impossible scores were removed from the smoothing.

Table 1

Summary of the Fit Measures for the Fitted Log-Linear Models in the Common-Item Nonequivalent Groups Design

	Test A			
Conditions	1 (Impossible scores included)		2 (Impossible scores excluded)	
Score distributions	New form	Old form	New form	Old form
N	1,404	1,735	1,404	1,735
DF	4,937	49,37	4,931	4,931
LR chi-square	2,071.95	2,936.71	1,083.96	1,086.15
AIC	2,095.95	2,960.71	1,107.96	1,110.15
	Test B			
N	223	433	223	433
DF	12,213	12,213	12,209	12,209
LR chi-square	1,215.51	1,814.88	1,159.22	1,722.84
AIC	1,231.51	1,830.88	1,175.22	1,738.84

Note. For Test A, the number of moments preserved for the total scores, anchor scores, and the cross-product (Total X Anchor) are 5, 5, 1, respectively. Fewer moments (i.e., 3, 3, 1) were preserved for Test B because of the relatively smaller sample size. DF = degrees of freedom, LR = likelihood ratio, AIC =Akaike information criterion.

Table 2

Summary Statistics for New and Old Forms (Tests A and B)

	Test A			
Score distributions	New form total	New form anchor	Old form anchor	Old form total
N	1,404	1,404	1,735	1,735
Mean	32.42	173.59	169.36	33.30
SD	3.27	17.15	19.32	3.77
Anchor/total correlation	0.50		0.61	
	Test B			
N	223	223	433	433
Mean	82.19	169.94	171.19	86.06
SD	20.50	18.84	20.92	21.51
Anchor/total correlation	0.79		0.86	

Finally, the FT deviates for Tests A and B were much larger (i.e., many FT deviates were larger than 3) in Condition 1 as compared to Condition 2, indicating a poor model fit for Condition 1. The larger FT deviates (for the most part) were associated with the zero frequencies for the impossible scores. For Test A, the FT deviates at the raw score points (in general) were smaller for Condition 2 than Condition 1, suggesting a better model fit when the impossible scores were removed from the smoothing (see Figure 1, which plots the FT deviates for conditions 1 and 2 for the new form anchor scores). Similarly, for Test B the FT deviates at the raw score points (in general) were smaller for Condition 2 than Condition 1, suggesting a better model fit when the impossible scores were removed from the smoothing (see Figure 2, which plots the FT deviates for conditions 1 and 2 for the new form anchor score). The same was true for the FT deviates for the old form anchor scores (i.e., they were relatively smaller in Condition 2 than Condition 1).

Equating Results for Tests A and B under Conditions 1 and 2

The means and standard deviations for the new and old forms for Tests A and B are given in Table 2. As seen in Table 2, the mean of the new form for Test A is slightly lower than the mean of the old form. The SD of the new form for Test A is also slightly smaller than the SD of the old form. The mean of the new form sample for Test AV is slightly larger than the mean of the old form sample for Test AV. The difference between the two means divided by the pooled SD of Test AV for the new and old form samples is 0.23, indicating that the new form sample is more able than the old form sample.

As seen in Table 2, the mean of the new form for Test B is slightly lower than the mean of the old form. The SD of the new form for Test B is also slightly smaller than the SD of the old form. The mean of the new form sample for Test BV is slightly smaller than the mean of the old form sample for Test BV. The difference between the two means divided by the pooled SD of Test AV for the new and old form samples is -0.06, indicating that the new form sample is about the same in ability as the old form sample.

The correlation of Test A with the MC anchor (Test AV) is moderate (0.5 to 0.61 for the new and old forms, respectively). The correlation of Test B with the MC anchor (Test BV) is fair (0.79 to 0.86 for the new and old forms, respectively). The actual differences (as evaluated using graphical and numerical approaches) in the equating results for Test A and B under Conditions 1 and 2 are presented next.

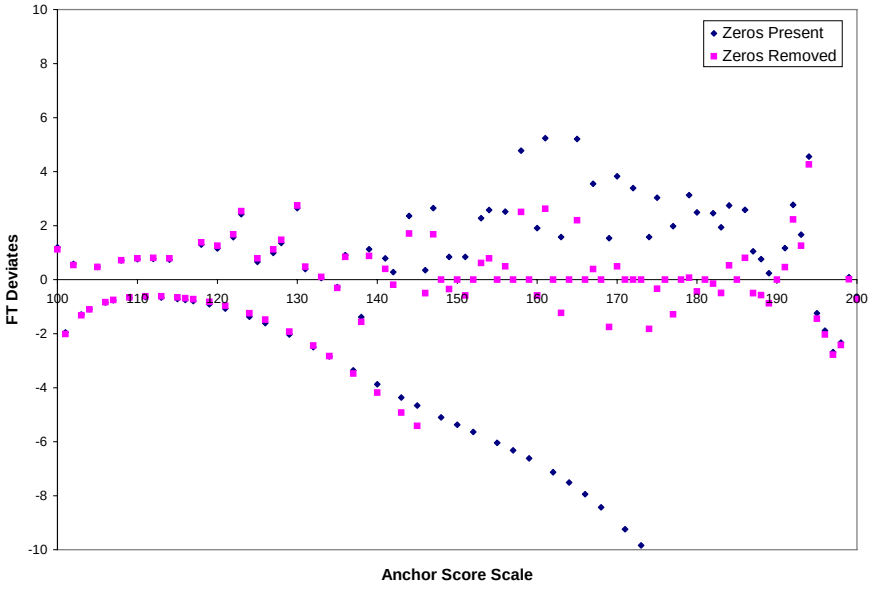

Figure 1. **Freeman-Tukey deviates for Test A anchor scores under Conditions 1 and 2.**

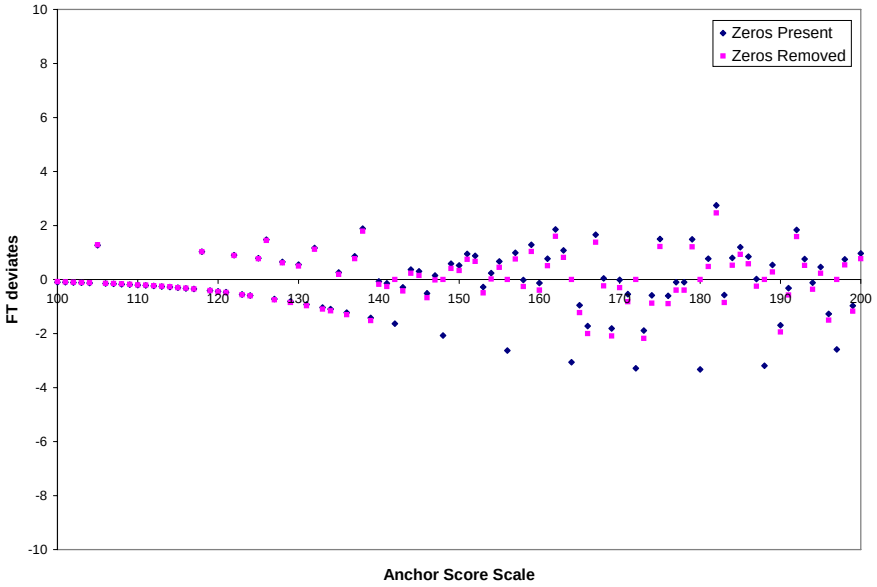

Figure 2. **Freeman-Tukey deviates for Test B anchor scores under Conditions 1 and 2.**

Difference Curve Under Conditions 1 and 2

As seen in Figure 3, the difference curve for CE equating for Test A shows a larger difference for the two equating functions derived under Conditions 1 and 2 at the tails of these functions as compared to the middle region, where the differences are smaller. However, the difference curve for FE equating shows a very small difference for the two equating functions for almost all parts of the score scale. The cut score range in scaled score units corresponds roughly to the 26 to 31 raw score region. As seen from the difference curve for the CE equating, there is a negligible difference (less than 0.5) for all score points in the cut score region except a raw score point of 29, where the difference is larger than 0.5. However, the difference curve for the FE equating indicates the difference is negligible for all parts of the cut score region and most parts beyond the cut score region.

As seen in Figure 4, the difference curve [2] for CE equating for Test B shows a larger difference for the two equating functions derived under Conditions 1 and 2 at the middle and higher score regions as compared to the lower score regions, where the differences are smaller. However, the difference curve for FE equating shows a very small difference for the two equating functions for almost all parts of the score scale. The cut score range in scaled score units corresponds roughly to the 59 to 103 raw score region. As seen from the difference curve for the CE equating, there is a non-negligible difference (greater than 0.5) for many score points in the cut score region. However, the difference curve for the FE equating indicates that the difference is negligible for all parts of the cut score region and most parts beyond the cut score region.

Results Using the RESD

The RESD between the two CE equating functions for Test A is 0.450.This is smaller than the DTM, indicating a potentially negligible difference between these equating functions. Similarly, the RESD between the two FE equating functions is 0.005, which is smaller than the DTM, indicating a potentially negligible difference between these equating functions.

The RESD between the two CE equating functions for Test B is 0.629, which is larger than the DTM, indicating a non-negligible difference between these equating functions. However, the RESD between the two FE equating functions is 0.02, which is smaller than the DTM, indicating a negligible difference between these equating functions.

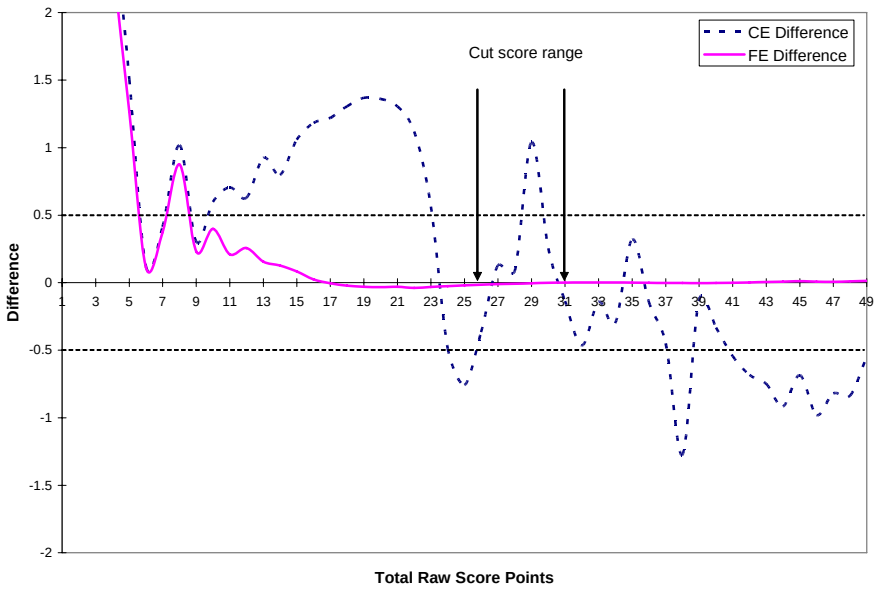

Figure 3. **Test A difference curve for chained equipercentile and frequency estimation equipercentile equatings derived under Conditions 1 and 2.**

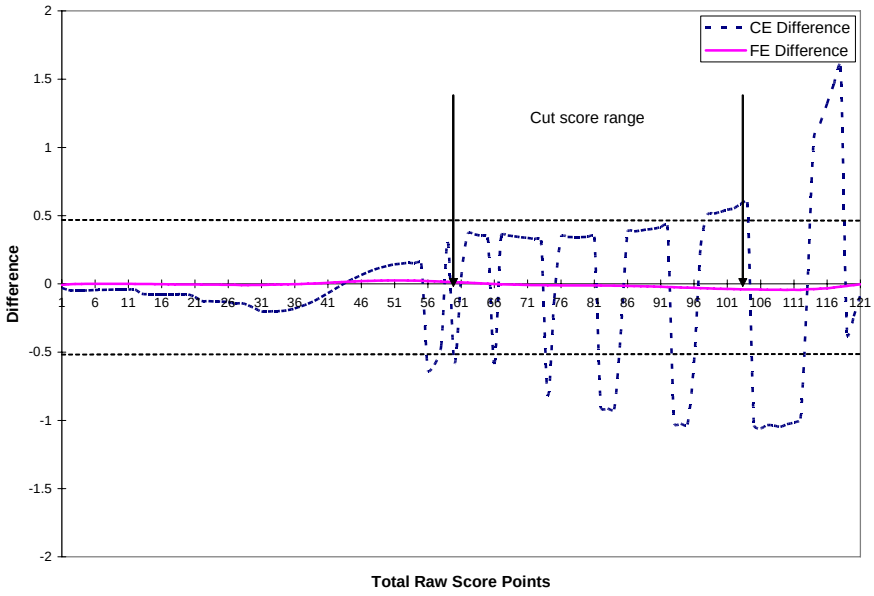

Figure 4. **Test B difference curve for chained equipercentile and frequency estimation equipercentile equatings derived under Conditions 1 and 2.**

Results Using the Conditional Standard Error of Equating

The CSEE for the Condition 1 and Condition 2 CE equatings [3] are show in Figures 5 and 6. As seen in these figures, the X axis does not extend to the full length of the score scale. Since there was sparse data at the tails of the distributions, thereby resulting in very large CSEE at these points, it was decided to include the CSEE that was 10 or smaller. As seen in Figure 5, for Test A the CSEE for Conditions 1 and 2 were similar for some score points, but not for other score points. Also, for score points where there were differences, a systematic pattern was not observed (i.e., the CSEE for Condition 2 were smaller than Condition 1 for some score points, but the reverse was true for other score points). Similar to Test A, the CSEE for Test B for Conditions 1 and 2 were similar for some score points, but not for other score points (see Figure 6). The irregular nature of the CSEE in the condition where the zeros were removed may be attributed to the impossible score points, which caused an unexpected drop from a high frequency to a zero frequency (i.e., for the impossible score) and back again to a high frequency, thereby making the standard error at these score points fluctuate drastically.

Passing Percentages

The actual pass percentages of test takers for different users with different cut scores are reported in Tables 3 and 4. As seen in Table 3, for Test A, the pass percentages for test takers using the CE equating function under Conditions 1 and 2 are identical for all users. Similarly, the pass percentages for test takers using the FE equating function under Conditions 1 and 2 are identical for all users. As seen in Table 4, for Test B, the pass percentages for test takers using the CE equating function under Conditions 1 and 2 are identical for most, but not all, users (although the difference in the pass rates for those users was very small, ranging from 0.4 to 1.79%). Similar to test A, the pass percentages for test takers using the FE equating function under Conditions 1 and 2 are identical for all users.

Discussion and Conclusions

The purpose of the present study was to evaluate the effect of impossible scores in an anchor score distribution on log-linear smoothing, and the final impact on equating results. Two conditions were studied (i.e., Condition 1, where the impossible zeros were included in the smoothing, versus Condition 2, where the impossible scores were excluded in the smoothing). The results of the smoothing in Conditions 1 and 2 were evaluated by using the likelihood ratio

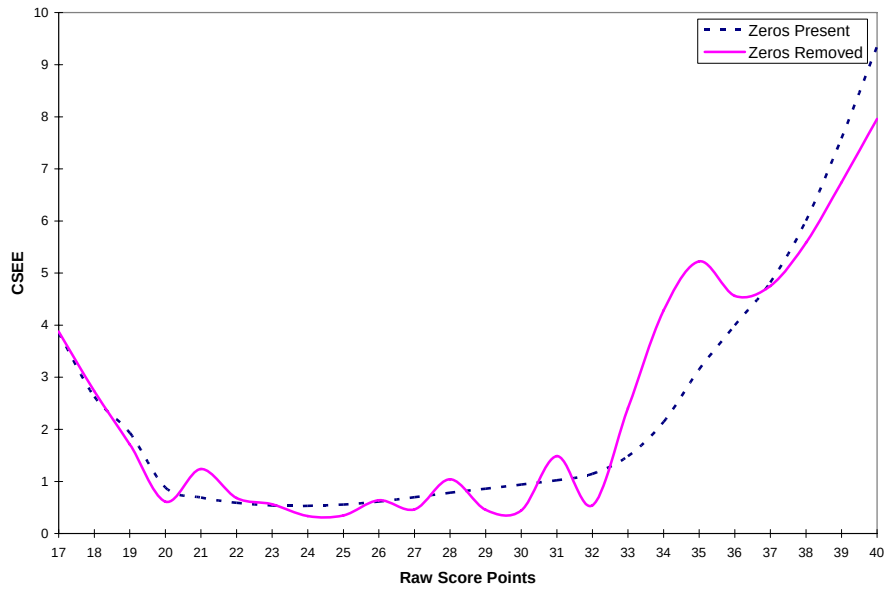

Figure 5. **Conditional standard errors of equating for chained equipercentile equating functions for Conditions 1 and 2 (Test A).**

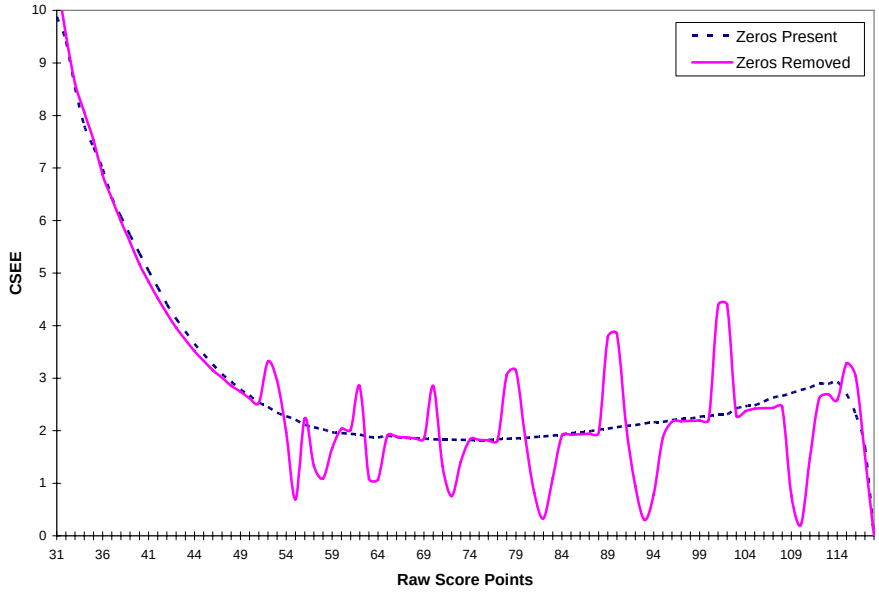

Figure 6. **Conditional standard errors of equating for chained equipercentile equating functions for Conditions 1 and 2 (Test B).**

Table 3

Actual Pass Percentages of Examinees Using Chained Equipercentile and Frequency Estimation Equipercentile Equating Conversions Derived From Conditions 1 and 2 (Test A)

User	CE 1	CE 2	FE 1	FE 2
1	95.66	95.66	95.66	95.66
2	95.66	95.66	94.44	94.44
3	84.83	84.83	84.83	84.83
4	84.83	84.83	76.42	76.42
5	76.42	76.42	76.42	76.42

Note. CE = chained equipercentile, FE = frequency estimation equipercentile.

Table 4

Actual Pass Percentages of Examinees Using Chained Equipercentile and Frequency Estimation Equipercentile Equating Conversions Derived From Conditions 1 and 2 (Test B)

User	CE 1	CE 2	FE 1	FE 2
1	87.89	87.89	82.96	82.96
2	63.68	63.68	63.68	63.68
3	**62.78**	**63.68 (2)**	62.78	62.78
4	62.33	62.33	62.33	62.33
5	57.40	57.40	57.40	57.40
6	54.71	54.71	54.71	54.71
7	**43.05**	**41.26 (4)**	46.19	46.19
8	**23.77**	**23.32 (1)**	25.56	25.56

Note. Bold indicates a difference in pass percentages. The numbers in parentheses indicate the actual number of examinees whose pass or fail status would change depending on which conversion was used. CE= chained equipercentile, FE= frequency estimation equipercentile. chi-square, AIC, and the FT deviates. The results of the equating in Conditions 1 and 2 were evaluated using the difference curve, RESD, CSEE, and actual passing rates.

The log-linear smoothing of the anchor score distributions as evaluated by the likelihood ratio chi-square and AIC indicated that the model fit for Condition 2 was much better than Condition 1 for both Tests A and B (i.e., the chi-square and AIC were smaller in Condition 2 than Condition 1, for both tests). The FT deviates scattered randomly around zero for both tests in Conditions 1 and 2. However, the FT deviates (in general) were smaller for Condition 2 as compared to Condition 1 for both tests, indicating a better model fit for Condition 2. As an additional evaluative criterion, the observed and fitted frequencies for Conditions 1 and 2 were

plotted on a graph (see Figures A1 and A2 for examples). The fitted frequencies in Condition 2 showed a much better fit to the observed data, as compared to Condition 1, where many fitted frequencies in the middle of the score distribution were pulled down because of the impossible scores. Findings from the smoothing indexes clearly indicated that the smoothing was better when the impossible scores were excluded in the smoothing. The next step was to examine whether the smoothing in both conditions impacted the final equated scores. If the actual equating does not differ much in Conditions 1 and 2, despite the apparent difference in the log-linear smoothing, the impact on test takers scores would be minimal. Findings from the equatings are summarized next.

The difference in the equating functions for the CE and FE equatings under conditions 1 and 2 were compared using graphical and numerical indexes. Findings for Test A are summarized as follows.

1. The difference curve between the two CE equating functions under Conditions 1 and 2 showed a potentially negligible difference for most parts of the cut score region. Similarly, the difference curve between the two FE equating functions under Conditions 1 and 2 showed a negligible difference for most parts of score scale and all cut score points.

2. The RESD obtained by comparing the CE equating functions from Conditions 1 and 2 was small (less than the DTM of 0.5), indicating a negligible difference between them. Similarly, the RESD results obtained by comparing the FE equating functions under Conditions 1 and 2 was zero (smaller than the DTM of 0.5), indicating a negligible difference between them.

3. The CSEEs for Conditions 1 and 2 did not show any systematic difference. The CSEE for Condition 1 was lower for some score points, while the CSEE for Condition 2 was lower at other score points

4. Finally, the pass percentages for the new form sample using the two CE equating functions showed that there were no differences in pass rates for all cut scores studied. Similarly, the pass percentages using the two FE equating functions showed that the pass rates were identical for all cut scores. These findings showed that although there were slight differences (especially for CE equating) that were observed between two

equating functions by using either a graphical or numerical approach, there was no effect of these differences on actual pass percentages.

Findings for Test B are summarized next:

1. The difference curve between the two CE equating functions under Conditions 1 and 2 showed a potentially negligible difference for the lower score regions and a potentially non-negligible difference for middle and higher score regions, including the cut score region. However, the difference curve between the two FE equating functions under Conditions 1 and 2 showed a potentially negligible difference for the total range of cut scores and the total score scale.

2. The RESD obtained by comparing the CE equating functions under Conditions 1 and 2 was large (greater than the DTM of 0.5), indicating a non-negligible difference between them. However, the RESD obtained by comparing the FE equating functions under Conditions 1 and 2 was zero, indicating a negligible difference between them.

3. The CSEEs for Conditions 1 and 2 did not show any systematic difference. The CSEE for Condition 1 was lower for some score points, while the CSEE for Condition 2 was lower at other score points.

4. Finally, the pass percentages for the new form sample using the two CE equating functions showed that there were no differences in pass rates for some cut scores (i.e., users), but not for others. However, the differences were quite small, ranging from 0.4 to 1.79%. [4] The pass percentages using the two FE equating functions showed that the pass rates were identical for all cut scores. These findings showed that although there were slight differences (especially for CE equating) that were observed between the two equating functions by using either a graphical or analytical approach, the effect of these differences on actual pass percentages was minimal.

An important question that should be addressed in a study comparing equating results from two different conditions is what can be done when a difference is found or when it is not found. Since pre-smoothing is considered an important step in equipercentile equating, it seems more appropriate to use Condition 2 (which results in a much better model fit) under all circumstances. However, smoothing is not the final product in equating; the equated scores are.

Therefore, if no difference is observed in the equated scores in either Condition 1 or 2, it should be a matter of little importance. A point to note is that although Condition 2 results in better model fit, it may be time-consuming to identify the impossible zeros and remove them in the smoothing under the tight operational time frame. For testing programs with strict timelines, usually a few days are allotted to conduct statistical analyses (i.e., item analysis, equating, etc) before test scores are reported to examinees and other users of test scores. In the case of Condition 2, it would add to the already short time frame to equate several new forms introduced in a testing administration. Therefore, the decision whether to choose Condition 1 or 2 (when results are similar for both conditions) should be evaluated in light of practical constraints, such as time required to conduct the work, reporting deadlines, etc.

However, if differences are noted in the equating functions in Conditions 1 and 2, then it is recommended to use the results from Condition 2, because the smooth frequencies used to equate scores in Condition 2 indicate a better fit to the observed data and are therefore more defensible, compared to the smoothed frequencies in Condition 1. In this context, the type of equating method used to equate scores may make a difference and the decision to use Condition 1 or Condition 2 may also be evaluated in light of these equating methods. For example, if the testing program uses classical linear equating methods such as Tucker or CE to equate test forms, then it would not matter if Condition 1 or Condition 2 was used. Smoothing does not impact the equated scores for these equating methods (although the CSEE may be impacted). In the case of equipercentile equating, results are expected to disagree when changes to the smoothing model are made, because the percentile ranks are influenced by the smoothed frequencies. In the current study, although differences were noted for the CE equated scores in Conditions 1 and 2; the FE equated scores were almost identical in both conditions. If future studies yield similar results, then one can be fairly confident that the presence or absence of the impossible zeros in the log-linear smoothing model may not impact the FE equated scores. Therefore if the testing program uses FE equating, then it would not matter whether the impossible zeros were present or removed in the smoothing

An interesting observation in this study was the difference in CE and FE equatings in Conditions 1 and 2 (for CE equatings a potentially non-negligible difference was observed at several score points, whereas for FE equatings there was almost no difference One possible reason why the CE results differed (even though the final impact on pass and fail was negligible) is that CE equating makes a direct link with the total test scores and the anchor test scores in two separate

single group equatings and then chains it with the new and old form scores to get the final equated scores. Therefore, any change in either the total frequencies or anchor score frequencies will have a direct impact on the CE results. However, the FE processes conditions on the anchor score. If the bivariate distribution is smooth, the conditional test score distribution for an impossible anchor score will be quite similar to the conditional test score distribution for a possible anchor score just one point higher or lower. Therefore, moving a large portion of the frequency at a possible anchor score to an impossible anchor score just one point higher or lower will have only a small effect on the distribution of the test scores. Therefore the FE equatings in Conditions 1 and 2 look very similar.

Limitations and Implications for Future Research

A limitation of the current study was the absence of a criterion equating function to which the equatings functions derived in Conditions 1 and 2 could be compared. If sample size for a particular test title is large enough, then a new form and an old form (already on scale) can be spiraled in one testing administration to create randomly equivalent groups. Then, the new form can be equated to the old form by employing the randomly equivalent groups design using either a linear or non-linear method and the results from that equating can be used as a criterion to compare the equating results from Conditions 1 and 2.

In this study, the correlation of the anchor test with the CR test was small to moderate (see Table 2). Future studies in this context may use other tests with different anchor-total correlations (i.e., very large or small). Results from such studies may be especially useful for further explaining the very small differences observed in the FE equating under Conditions 1 and 2. According to Livingston (2004), in case of FE or Tucker equating , when the correlations between the anchor and total test scores depart from 1.00, these equating methods adjust differences in test form difficulty as if the equating samples are more similar than the anchor scores indicate. In the current study where the anchor scores were smoothed under two different conditions, such smoothing may have more or less impact on the actual equating, depending on the anchor-total correlation. If the correlation is low, then, following on Livingston's idea, the equatings under Condition 1 or 2 may not differ much.

This study examined two tests that had a score scale that extended over 100 points. Since there were over 110 raw score points, the raw and scaled score correspondence (after taking into account chance level) roughly fell in the 1: 1.2 ratio. The incidence of impossible scores would

probably increase for tests where there is a bigger difference in this ratio. Therefore, although in this study the equating results were quite similar for Conditions 1 and 2, it is possible that for other tests where the scaled score range is much larger than the raw score range, there will be many more impossible score points and the effect of removing them in the smoothing may have a more drastic impact on the equating. Therefore, future studies can be conducted using such tests (with relatively larger scale score range than raw score range) to evaluate the impact of removing the impossible scores from smoothing on the equating results.

Finally, the current study involved only two replications (the smoothing and equating analyses) with two different tests. Although repeating the analyses with data from two tests provides some cross validation for these results, repeating the analyses with more tests and/or with several random samples selected with replacement from the current test data is desirable, to allow for a stronger generalization of these results.

The study was important as it evaluated whether different ways of smoothing a score distribution having impossible scores lead to different smoothing and equating results. Since in various testing programs new CR forms are often put on scale through an external MC anchor, the presence of these impossible scores is a constant concern. As the tests used in the current study have cut scores, the actual pass/fail decisions based on the different equating functions from Conditions 1 and 2 were also reported. As is evident, the findings of this study suggest that although there were slight differences in the equated scores (especially for CE equating) in Conditions 1 and 2, the effect of this difference on actual pass and fail status was minimal. However, this could change if new cut scores are introduced as results may look somewhat different. Also, in a test which does not use cut scores, assessing the differences (if any) in most parts of the score scale and not just the cut score region is important.

References

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Proceedings of the symposium on applications of statistics* (pp. 27–47). Amsterdam: North-Holland.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. 94-10). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Program Statistics Research Tech. Rep. No. 87-79). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

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

Livingston, S. A. (1996). [Review of the book *Test Equating*]. *Journal of Educational Measurement, 33*(3), 369–373.

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

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

Notes

[1] Although impossible scores are often observed in scaled score distributions, they can also occur with raw scores. For example, when the raw score is the sum of the scores from two raters, where raters are not allowed to disagree, then the only possible scores are even numbers. Also note that if a test being equated has a large sample size and the new and old form anchor tests are the same, using raw anchor scores in the equating may be preferable, as this would eliminate the problem of impossible zeros that arise from using scale anchor scores.

[2] A systematic but unusual pattern is observed in the difference curve where the difference is greater than zero for some scores followed by a difference that is lower than zero for other scores, and so on. A partial explanation for this occurrence is that in the zeros-excluded condition it is possible for the two adjacent anchor scores to have the same percentage of test takers below the two score points. For example, if Score Point 34 in the reference form anchor is impossible, but Score Point 35 is possible, then in the zero excluded condition the percentage of test takers below these two score points will be the same. Therefore, if two adjacent scores in the new form convert to two adjacent scores (like the ones mentioned above), the two adjacent scores in the new forms will get the same converted scores. If there are several impossible points like this in the reference anchor score, the conversion line will look somewhat irregular (i.e., the conversion line will rise monotonically for some raw scores, remain relatively constant for subsequent raw scores, and then rise again). This irregularity in the conversion line in the zeros-excluded condition results in the irregularity in the difference curve. Although for this test the irregularity in the conversion line was caused by excluding the zeros, it may not affect the difference curve for another test in exactly the same way. For example, if the percentage of test takers below two adjacent scores in the new form are far apart (e.g., percentage below for Score Points 34 and 35 are 62 and 74), then it is less likely that their corresponding anchor score points in the reference form will be adjacent (recall from above that two adjacent score points in the reference form anchor can have the same percentage of test taker below in the zeros-excluded condition) and therefore will convert to different scores. This may be the reason why the difference curve in Figure 3 is not as irregular as in Figure 4.

[3] Since the difference between the FE equating functions from Conditions 1 and 2 was close to zero for most score points (small enough not to make a difference in pass rates), it seemed unnecessary to evaluate the difference using the CSEE approach.

[4] Although the difference in pass rates was quite small, the actual impact of test takers may vary depending on the actual sample size (a large sample may affect many test takers), difference between the mean of the test taking group and the cut score (a larger difference usually leads to higher consistency of classification as pass or fail), etc.

Appendix

Observed and Fitted Frequencies for New Form Anchor Scores

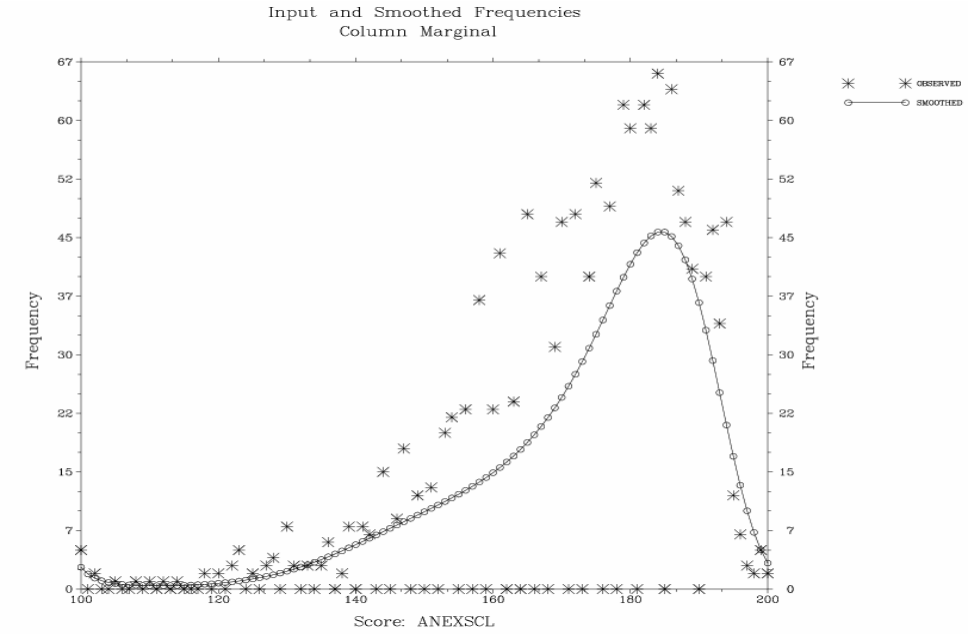

Figure A1. **Observed and fitted frequencies for new form anchor scores in Test A (Condition 1)**

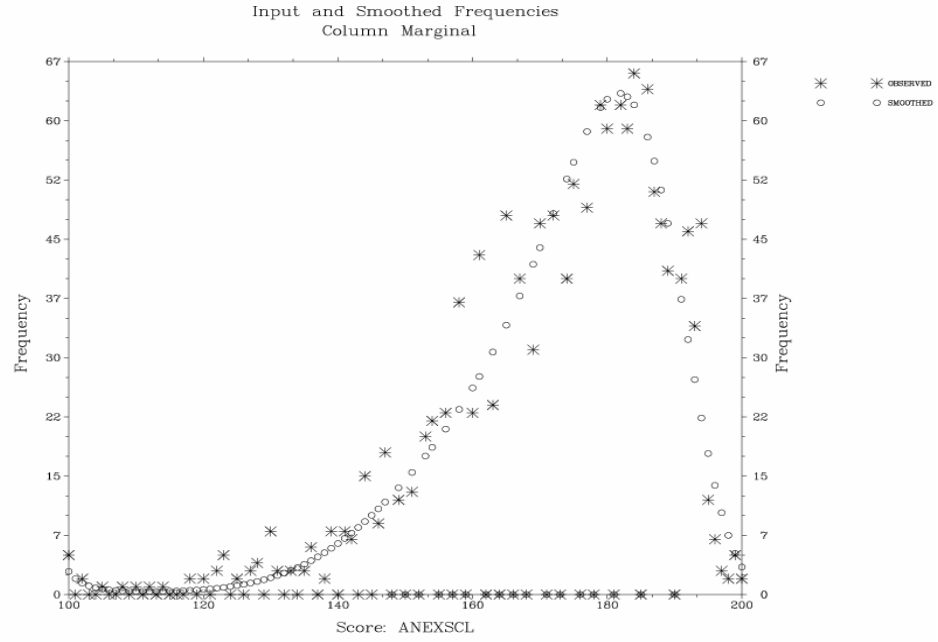

Figure A2. **Observed and fitted frequencies for new form anchor scores in Test A (Condition 2)**