

*Chained Versus Post-Stratification
Equating in a Linear Context:
An Evaluation Using Empirical Data*

Gautam Puhan

February 2010

ETS RR-10-06



**Chained Versus Post-Stratification Equating in a Linear Context:
An Evaluation Using Empirical Data**

Gautam Puhan
ETS, Princeton, New Jersey

February 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

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

Technical Review Editor: Dan Eignor

Technical Reviewers: Insu Paek and Mary Grant

Copyright © 2010 by Educational Testing Service. All rights reserved.

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

SAT and PSAT/NMSQT are registered trademarks of the
College Board.



Abstract

This study used real data to construct testing conditions for comparing results of chained linear, Tucker, and Levine-observed score equatings. The comparisons were made under conditions where the new- and old-form samples were similar in ability and when they differed in ability. The length of the anchor test was also varied to enable examination of its effect on the three different equating methods. Two tests were used in the study, and the three equating methods were compared to a criterion equating to obtain estimates of random equating error, bias, and root mean squared error (RMSE). Results showed that for most of the conditions studied, chained linear score equating produced fairly good equating results in terms of low bias and RMSE. In some conditions, Levine-observed score equating also produced low bias and RMSE. Although the Tucker method always produced the lowest random equating error, it produced a larger bias and RMSE than either of the other equating methods. Based on these results, it is recommended that either chained linear or Levine score equating be used when new- and old-form samples differ in ability and/or when the anchor-to-total correlation is not very high.

Key words: Tucker equating, Levine equating, chained linear equating, equating error, equating bias

Table of Contents

	Page
Introduction.....	1
Previous Studies Comparing Linear and Nonlinear PSE and CE Methods.....	2
Method	3
Design Used for Test X.....	3
Design Used for Test Y.....	6
Procedure Followed to Evaluate the Equating Methods Under the Different Conditions.....	7
Results.....	8
Results for Test X	9
Results for Test Y	20
Discussion and Conclusion.....	25
Why Did Tucker Equating Perform Poorly? A Partial Explanation.....	26
Can Tucker Equating Perform Better Than Chained Equating in Some Conditions?	28
Limitations and Future Research	29
References.....	31
Appendix.....	33

List of Tables

	Page
Table 1. Summary Statistics for New-Form, Old-Form, and Anchor Tests in the Full Sample (Test X)	11
Table 2. Summary Statistics for New-Form, Old-Form, and Anchor Tests in the Full Sample (Test Y)	12
Table 3. Average SEE and Bias for Test X for the Small, Moderate, and Large Ability Difference Conditions	13
Table 4. Average Standard Error of Equating (SEE) and Bias for Test Y for the Moderate and Moderately Large Ability Difference Conditions	21

List of Figures

	Page
Figure 1. Graph showing two alternate subforms (Form X1 and Form X2) created from one original Form X.	4
Figure 2. Conditional standard error of equating (CSEE) for small ability difference condition (Test X).	14
Figure 3. Conditional standard error of equating (CSEE) for moderate ability difference condition (Test X).	15
Figure 4. Conditional standard error of equating (CSEE) for large ability difference condition (Test X).	16
Figure 5. Conditional bias (CBias) for small ability difference condition (Test X).	17
Figure 6. Conditional bias (CBias) for moderately large ability difference condition (Test X).	18
Figure 7. Conditional standard error of equating (CSEE) for large ability difference condition (Test X).	19
Figure 8. Conditional standard error of equating (CSEE) and conditional bias (CBias) for moderate ability difference condition (Test Y).	23
Figure 9. Conditional standard error of equating (CSEE) and conditional bias (CBias) for moderately large ability difference condition (Test Y).	24

Introduction

When parallel forms of the same test are administered to nonequivalent groups of examinees, then a third test (i.e., common or anchor test) is often used as a yardstick for conducting the equating. Suppose Form A is given to Group 1 along with Anchor Test X, and Form B is given to Group 2 along with Anchor Test X, then the performance of each group on the anchor test provides information that is useful for comparing the scores on both forms of the test. This equating design is known as the *common item nonequivalent groups design* (Kolen & Brennan, 2004).

There are two widely used equating methods under the common item nonequivalent groups design (i.e., the post-stratification equating method, or PSE, and chained equating method, or CE). If Form A and Form B are equated, then in PSE linear equating (i.e., Tucker equating), the mean and standard deviations of Form A and Form B for a common synthetic population are estimated and then the equating is conducted by substituting these estimates into the basic formula for linear equating. In chained linear equating (Dorans, 1990; Kolen & Brennan, 2004), Form A scores are equated to scores on the anchor test using Group 1 (e.g., a score on Form A and the anchor test are considered equivalent if they are the same distance in standard deviation units above or below their respective mean scores) and Form B scores are equated to scores on the anchor test using Group 2. These two conversions are then chained together to produce the transformation of Form A scores to Form B scores. Both PSE and CE can also be used in an equipercentile framework, and the equipercentile versions of the PSE and CE methods are commonly known as *frequency estimation* and *chained equipercentile equating* (Kolen & Brennan, 2004; Livingston, 2004).

In actual testing programs, equating is often conducted under varied conditions (e.g., equating when new- and old-form samples differ in ability, equating when the correlation between the anchor test and the total test is not optimal, etc.). Some important questions arise in this context. Do PSE and CE linear score equating methods produce similar results when new- and old-form samples are similar in ability? Do PSE and CE linear score equating methods produce similar results when new- and old-form samples are not similar in ability? Does length of the anchor test (which may affect the correlation between the total test and anchor test) influence linear PSE and CE methods differently? The first two questions have been examined in earlier research (see Livingston, Dorans, & Wright, 1990; Sinharay & Holland, 2007; Wang,

Lee, Brennan, & Kolen, 2006). The third question has not received much attention, except in the study by Wang et al. (2006).

The purpose of this study is to evaluate the performance of linear PSE (i.e., Tucker) and chained linear score equating methods under conditions when new- and old-form samples are similar in ability, and when they are not similar in ability. Furthermore, within the ability difference conditions, the length of the anchor test was also varied to examine its effect on these equating methods. The Levine-observed score equating method, simply called *Levine equating* throughout this paper, was also be compared with the other two equating methods because it is believed to offer the same benefits as PSE (i.e., strong theoretical standing) but without the bias of PSE (Livingston, 2004). This method uses estimates of true scores in determining the mean and standard deviations of the new and old forms. Then the equating is conducted by substituting these estimates into the basic formula for linear equating (see Kolen & Brennan, 2004, pp. 109–115 for details). It seemed reasonable to include only the Levine equating method and not the Levine-true score equating method in this study because the other methods included in the comparison (i.e., Tucker and chained linear) are also observed score equating methods.

Note that although a similar comparison can be made between the nonlinear PSE and CE methods, which are often the methods of choice when new- and old-form samples differ in difficulty, this study focused on linear methods (see Harris & Kolen, 1990; Sinharay & Holland, 2007, and Wang et al., 2006 for a comparison of PSE and CE methods in a nonlinear context).

Previous Studies Comparing Linear and Nonlinear PSE and CE Methods

Equating studies by Marco, Petersen, and Stewart (1983) and Sinharay and Holland (2007) conducted in nonlinear contexts, and Livingston, Dorans, and Wright (1990) and Wang et al. (2006) conducted in both linear and nonlinear contexts, suggested that CE methods tend to produce less equating bias than produced by PSE methods when groups differ in ability. Harris and Kolen (1990) compared chained equipercentile and frequency estimation equatings and found that these methods produced different results, especially when new- and old-form samples differed in ability. They suggested using PSE methods because these methods have a better theoretical standing in comparison to CE methods. However, Marco et al. (1983) and Livingston, et al. (1990) supported the use of CE methods in situations where there was a large ability difference in the groups that took both forms of the test.

According to Livingston (2004), when groups differ in ability and the correlation between the total test and anchor test is not high, the PSE method adjusts form difficulty in a way that treats the two groups as if they were more similar than they actually are, thereby leading to a biased equating. On the other hand, the CE method uses a scaling procedure, which is symmetric in nature, and is not much influenced by the size of the correlation between the anchor test and the total test. This results in a less biased equating for CE when groups differ in ability.

Bias is one component of equating error. The other component is random error, which occurs whenever a sample, instead of the population, is used to derive the equating function. Sinharay and Holland (2007) found that although the bias is slightly higher for PSE compared to the CE method, random equating error is slightly lower for PSE compared to the CE method. A similar finding was observed by Wang et al. (2006). Random equating error, in addition to bias, should also be considered when deciding which equating method to use when groups differ in ability. For example, if an equating method is only slightly more biased in comparison to another method but produces a much smaller random equating error compared to the other method, then it may be desirable to use the method with the much smaller random equating error.

Method

Data for this study consisted of examinee responses from two tests (referred to as Test X and Test Y, respectively). The purpose was to compare the different equating functions derived for two pseudo forms of Test X (referred to as Forms X1 and X2) and from one form of Test Y (referred to as Form Y1), which was equated to itself (i.e., Form Y1). Equating a test to itself helped emulate a situation where the new and old forms are very similar in difficulty. It also made it possible to use the identity function as a natural criterion. The conditions under which the different equating methods were compared were slightly different for the two tests. Therefore a brief description of the design used for each test is presented in the following section.

Design Used for Test X

Forms X1 and X2 are pseudo forms created from one single form known as Form X (see Figure 1 for illustration). There were 23,418 examinee responses available for Form X. As seen in Figure 1, Form X (consisting of 120 items) was divided into two alternate subforms, Form X1 and Form X2, each consisting of 84 items. The shaded portion indicates the common section of

48 items between Form X1 and Form X2. This procedure of using one original form to build two subforms and conducting the equating on the subforms allowed for the creation of a strong equating criterion, which is discussed in the following section.

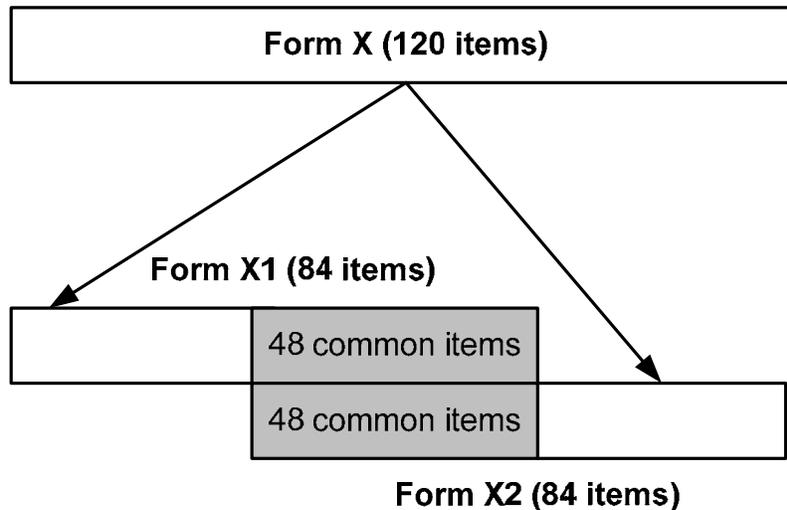


Figure 1. Graph showing two alternate subforms (Form X1 and Form X2) created from one original Form X.

Equating criterion. When comparing different methods to evaluate which equating method performs better, it is necessary to compare the results of each method to a criterion (i.e., a close proxy of the equating function in the population). The design used for Test X facilitated the computing of a criterion equating function. Essentially, since Forms X1 and X2 were created from one Form X (taken by 23,418 examinees), all the examinees took Form X1 and also Form X2. Therefore, Form X1 can be equated directly to Form X2 using a single group (SG) equating design. By using the SG design with such a large data set, one can be fairly confident that the resulting conversion is a very good approximation of the equating in the population (see Livingston, 1993, who first proposed this method for deriving a criterion equating function).

Identifying strong and weak test-taking groups. Since the purpose of the study was to evaluate the effect of ability differences between the new- and old-form samples on different equating methods, Forms X1 and X2 were assigned to examinee groups that differed in ability. A natural way to partition examinees into different ability groups would be to do so based on their scores on the anchor test (i.e., the 48 common items between Form X1 and Form X2). However,

this method has been shown to favor PSE equating methods (see Livingston, Dorans, & Wright, 1990) and was therefore not used in this study. To replicate actual differences that are observed in real testing administrations, the assignment of examinees to Form X1 or Form X2 was based on actual ability differences observed in real test administrations. As mentioned earlier, Form X had examinee responses from 23,418 examinees. These were accumulated examinee responses from four different test administrations in which Form X was administered. Since Form X was administered (without any changes) in these four different test administrations, any difference in mean scores across these four administrations would be solely due to ability differences between the four groups taking Form X. This information regarding the mean score was used to identify examinee groups of differing ability.

Conditions based on ability difference and length of anchor test. For Test X, three ability difference conditions were examined. Two out of the four test administration groups mentioned earlier, whose mean scores showed the smallest difference (standardized mean difference, or SMD, on the 48 anchor items was 0.03), were assigned to Form X1 and Form X2. This condition was referred to as the *small ability difference* condition. For the second condition, two out of the four groups whose mean scores showed the maximum difference (SMD was 0.19) were assigned to Forms X1 and X2. This condition was referred to as the *moderately large ability difference* condition. Although an SMD of 0.19 is indicative of a fairly large ability difference between the new- and old-form samples, there are instances in actual testing programs when new- and old-form samples differ even more. However, such a large difference between the four test administrations was not observed in the current data. To simulate a third condition, where the ability difference was larger than what was examined in the second condition, the new- and old-form samples from the second condition were used but with a slight modification. The new-form sample was unaltered and was assigned as the new-form sample in the third condition. However, a slightly more able sub-sample from the old-form sample was selected and assigned to the old-form in the third condition, referred to as the *large ability difference* condition (SMD was 0.28). The sub-sampling was done by using examinee ethnicity as a stratifying variable, and fewer non-White examinees (who were, on average, less able than White examinees) were selected in the sub-sample, resulting in a higher mean score than for the full sample. Within each ability difference condition (i.e., small, moderately large, and large ability difference conditions), the number of anchor items used to conduct the common item

nonequivalent groups equating was also varied. The equatings within each ability difference condition used 48, 36, 24, or 12 anchor items. The length of the anchor test was expected to result in different anchor-total test correlations, which may affect PSE and CE equatings differently. Therefore, 12 conditions were examined (3 ability difference conditions \times 4 anchor test length conditions).

Design Used for Test Y

The design followed for Test Y was to equate one form of this test (i.e., Form Y1) to itself, using examinee groups of differing ability. Form Y1 consisted of 50 items and was an operational form used in actual test administrations. This facilitated using a group of examinees (e.g., strong group) taking Form Y1 in a particular test administration as the new-form sample and another group of examinees (e.g., weak group), which had also taken Form Y1 in another test administration as the old-form sample and then equating Form Y1 to itself using these two different examinee groups.

Equating criterion. Since Form Y1 was equated to itself, the identity function (i.e., the original score scale of Form Y1) was used as the criterion, and the results of the different equating methods were compared to this criterion. Formally, the identity function can be defined as placing the scores of the new form onto the scale of the old form using a direct linear equating with a slope of 1 and intercept of 0.

Identifying strong and weak test-taking groups. As mentioned earlier, Form Y1 was an operational form used in actual test administrations. Since the same form was administered in four different administrations (similar to Test X, discussed earlier), any change in examinee performance across different test administrations can be attributed solely to differences in ability. Therefore, the mean score of the examinees in these different test administrations was used to identify groups of differing abilities.

Conditions based on ability difference and length of anchor test. For Test Y, two ability difference conditions were examined. Two out of the four test administration groups whose mean scores showed a moderate difference (SMD on the 50 items was 0.12) were designated as new- and old-form samples. This condition was referred to as the *moderate ability difference* condition. For the second condition, two out of the four groups whose mean scores showed the maximum difference (SMD was 0.21) were designated as the new- and old-form samples. This condition was referred to as the *moderately large ability difference* condition. The

number of anchor items used to conduct a common item nonequivalent groups equating was also varied. The equatings within the moderate and moderately large ability difference condition used 20 and 12 anchors items, respectively. Therefore, four conditions were examined (two ability difference conditions \times two anchor test length conditions).

Procedure Followed to Evaluate the Equating Methods Under the Different Conditions

The chained linear, Tucker, and Levine score equating methods were compared within each ability and anchor test length condition. After the ability difference and anchor test length conditions were determined for each test, the study was conducted as follows. Although the steps described are specific to Test X, the same procedure was followed for Test Y.

Step 1. To estimate the criterion equating function, Form X1 was equated to Form X2 (i.e., setting the means and standard deviations equal in the two forms) using the SG equating design with the total data ($N = 23,418$). The resulting linear conversion was considered the criterion to which the chained linear, Tucker, and Levine equatings were compared. Although a direct equipercentile equating, instead of direct linear equating, in the single group could be used as a criterion, it was not used in order to avoid any confounding that may arise because of comparing linear equating methods with a nonlinear equating criterion (see Kim, von Davier, & Haberman, 2006, who provided a rationale for using a linear equating criterion in a similar context). For Test Y, this step was not needed because the identity equating served as the criterion equating function.

Step 2. For a particular condition (e.g., large ability difference condition using anchor test length of 48), sample sizes of 1,000 each were drawn with replacements from the Form X1 and Form X2 data sets. As explained earlier, the samples assigned to these two forms differed in ability (i.e., the less able group was assigned to Form X1 and the more able group was assigned to Form X2). Then three score equatings (chained linear, Tucker, and Levine) were conducted to equate Form X1 to Form X2 using the common item equating design.

Step 3. Step 2 was repeated 500 times and, based on the repeated samples' conditional standard error of equating (CSEE), the average of the CSEE (AvgCSEE), conditional bias (CBias), and root mean squared bias (Bias^2) were computed and compared for the nonequivalent groups anchor test (NEAT) and single group with nearly equivalent test forms (SiGNET) equatings. The root mean squared was computed to prevent large positive and negative differences from cancelling out each other. When computing the AvgCSEE and bias, the CSEEs

and CBias values were appropriately weighted using the raw proportion of examinees at each score point in the total population data. The root mean squared deviation (RMSD) was also calculated at the total test score level for the NEAT and SiGNET equatings. The formula is

$$\text{RMSE}=\sqrt{\text{Bias}^2+\text{AvgSEE}^2}$$

where Bias^2 is the sum of the squared conditional bias values weighted by the raw proportion of examinees at each score point. The RMSD is a useful statistic because it provides an estimate based on combining information from random and systematic error. A detailed description of these statistical indices is provided in Kim, Livingston, and Lewis (2009).

Although the different equating methods were compared relative to each other, the practical criterion of the *difference that matters*, or DTM, (Dorans & Feigenbaum, 1994) was also used to evaluate the different equating methods. The DTM is a unit equal to one-half of a reporting unit. Since for the tests used in this study, scores progressed in 1-point increments, the DTM was defined as any score difference that is equal to or greater than 0.5. Using a DTM criterion seemed reasonable because if a difference existed between the variability and accuracy indices obtained using the three equating methods, but the actual values were smaller than the DTM, then the differences are probably ignorable because they may not result in a practical difference in the examinees' reported scores.

Results

The summary statistics for Tests X and Y under the different ability difference and anchor test length conditions are presented in Tables 1 and 2, respectively. For Test X, the standardized mean difference, or SMD, which is the difference in the new- and old-form sample means on the anchor items divided by the pooled standard deviation on the anchor items, was 0.03 for the small ability difference condition, 0.19 for the moderately large ability difference condition, and 0.28 for the large ability difference condition (see Table 1). As seen in Table 1, the anchor-to-total test correlation was fairly high for the 48- and 36-item anchor conditions (min. = 0.908 and max. = 0.957 across the three ability difference conditions) and moderate to moderately high for the 24- and 12-item anchor conditions (min. = 0.774 and max. = 0.870). As seen in the small ability difference condition, the mean total score on the new form was lower ($\bar{X} = 55.046$) than the mean total score on the old form ($\bar{X} = 57.996$). Considering that the

difference in ability between new- and old-form samples in this condition was very small, one can infer based on these mean scores that the new form was more difficult than the old form.

For Test Y, the SMD was 0.12 for the moderate ability difference condition and 0.21 for the moderately large ability difference condition (see Table 2). As seen in Table 2, the anchor-to-total test correlations were fairly high for the 20-item anchor condition (min. = 0.887 and max. = 0.904 across the two ability difference conditions) and moderately high for the 12-item anchor item condition (min. = 0.792 and max. = 0.813).

The overall accuracy and variability of equatings for the three linear score equating methods under different ability difference and anchor test length conditions were estimated using the average SEE, bias, and average RMSD indices. For Tests X and Y, these results are presented in Tables 3 and 4, respectively. The conditional standard error of equatings or CSEEs for the small, moderate, and large ability difference conditions for Test X are presented in Figures 2, 3, and 4, respectively. The conditional bias for the small, moderate, and large ability difference conditions for Test X are presented in Figures 5, 6, and 7, respectively. The CSEEs and conditional bias for the moderate and moderately large ability difference conditions for Test Y are presented in Figures 8 and 9, respectively.

Results for Test X

Average SEE, bias, and RMSD results. As seen in Table 3, for the small ability difference condition, the average SEE for the Tucker method was always smaller than the average SEE for the chained linear or Levine methods. The average SEE for the chained linear method was always smaller than the average SEE for the Levine method. This was true for all anchor test length and ability difference conditions (similar findings have been reported in Sinharay & Holland, 2007). The average SEEs for the three equating methods were largest for the 12-item anchor condition and became progressively smaller for the larger anchor conditions (i.e., 24, 36, and 48 anchor items). This was true for all three ability difference conditions. The bias for the Levine equating method was smaller than the Tucker and chained linear methods for all anchor lengths in the small ability difference condition. However, for the moderate and large ability difference conditions, the chained linear method produced the smallest bias in some conditions, while the Levine method produced the smallest bias in other conditions. Unlike the average SEE, which decreased as length of anchor items increased, bias did not follow such a predictable pattern, especially for the chained linear and Levine equatings. The bias for the

Tucker equating method followed a somewhat predictable pattern and increased as the number of items in the anchor test decreased in the moderate and large ability difference conditions. The RMSE, which provides a useful summary of total error by combining the random and systematic equating error, showed that the chained linear method produced the smallest RMSE for most of the studied conditions. Finally, in terms of bias or RMSE, the Tucker method performed the worst in all three ability difference conditions.

Conditional standard errors and bias. Although the average statistics described in the previous section provide a useful summary of random and systematic equating error, the CSEEs and conditional bias values are often considered more informative because they indicate the amount of variability and accuracy at each score point. Since CSEEs and conditional bias values tend to be less stable at score points with less data, it was decided to focus on the score points between the 5th and 95th percentiles because most of the data was observed within this score range. For the new form, this score roughly ranged between score points 35 and 72 for the three ability conditions and was used to evaluate the CSEEs and conditional bias values for Test X. As seen in Figure 2, for the small ability difference condition, the CSEEs of the Tucker method were smaller than the CSEEs of the chained linear or Levine methods, and the CSEEs of the chained linear method were lower than the CSEEs of the Levine method. This trend was also observed for the moderate and large ability difference conditions. Also, as seen in Figures 2–4, when anchor size was held constant, the CSEEs did not differ much across the three different ability difference conditions (e.g., the CSEEs for the 12-anchor small ability difference condition were very similar to the CSEEs for the 12-anchor moderate or large ability difference conditions). As seen in Figures 2–4, the CSEEs are smaller than the DTM for the 35-to-72 score range (represented by the arrows) for the 48- and 36-anchor items condition and begins to fall outside the DTM range for the 24-anchor items conditions, and even more so for the 12-anchor items condition. Similar to the average statistics described earlier, the CSEEs become progressively smaller as the sample size of the anchor items increase from 12 to 48, thus serving as an important reminder that random equating error is not only dependent on the sample size of test takers but also is dependent on the number of anchor items.

Table 1**Summary Statistics for New-Form, Old-Form, and Anchor Tests in the Full Sample (Test X)**

Score distributions	NF total	NF anchor	NF anchor	NF anchor	NF anchor	OF total	OF anchor	OF anchor	OF anchor	OF anchor
# of items	84	48	36	24	12	84	48	36	24	12
Small ability difference condition (SMD = 0.03)										
<i>N</i>	6,580					6,798				
Mean	55.046	33.391	25.321	16.973	8.410	57.996	33.174	25.178	16.859	8.303
<i>SD</i>	10.567	6.164	4.618	3.215	2.000	9.894	6.067	4.540	3.171	1.990
Reliability	0.863	0.777	0.708	0.634	0.531	0.866	0.792	0.734	0.662	0.530
Anchor/total correlation		0.949	0.908	0.854	0.775		0.956	0.920	0.870	0.777
Moderately large ability difference condition (SMD = 0.19)										
<i>N</i>	6,469					6,580				
Mean	52.779	32.156	24.375	16.358	8.072	58.426	33.391	25.321	16.973	8.410
<i>SD</i>	11.147	6.518	4.903	3.393	2.117	10.013	6.164	4.618	3.215	2.000
Reliability	0.872	0.792	0.736	0.662	0.548	0.868	0.795	0.736	0.663	0.550
Anchor/total correlation		0.953	0.918	0.867	0.785		0.957	0.921	0.870	0.788
Large ability difference condition (SMD = 0.28)										
<i>N</i>	6,469					5,449				
Mean	52.779	32.156	24.375	16.358	8.072	59.261	33.884	25.701	17.224	8.546
<i>SD</i>	11.147	6.518	4.903	3.393	2.117	9.461	5.844	4.374	3.071	1.936
Reliability	0.872	0.792	0.736	0.662	0.548	0.855	0.776	0.706	0.634	0.523
Anchor/total correlation		0.953	0.918	0.867	0.785		0.953	0.911	0.858	0.774

Note. Total single group (SG) data ($N = 23,418$); NF mean and $SD = 53.944$ and 10.730 , respectively; and OF mean and $SD = 57.504$ and 10.256 , respectively. NF = new form, OF = old form, SMD = standardized mean difference.

Table 2**Summary Statistics for New-Form, Old-Form, and Anchor Tests in the Full Sample (Test Y)**

Score distributions	NF total	NF anchor	NF anchor	OF total	OF anchor	OF anchor
# of items	49	20	12	49	20	12
Moderate ability difference condition (SMD = 0.12)						
<i>N</i>		<u>1,140</u>			<u>2,067</u>	
Mean	24.139	9.160	5.059	25.038	9.169	5.271
<i>SD</i>	7.176	3.364	2.208	7.399	3.235	2.263
Reliability	0.807	0.634	0.507	0.853	0.697	0.544
Anchor/total correlation		0.887	0.792		0.904	0.811
Moderately large ability difference condition (SMD = 0.21)						
<i>N</i>		<u>1,140</u>			<u>1,695</u>	
Mean	24.139	9.160	5.059	25.606	9.753	5.418
<i>SD</i>	7.176	3.364	2.208	6.678	3.244	2.258
Reliability	0.807	0.634	0.507	0.821	0.668	0.532
Anchor/total correlation		0.887	0.792		0.902	0.813

Note. Total single groups (SG) data ($N = 6, 418$); NF and OF mean and $SD = 24.669$ and 7.283 , respectively.

NF = new form, OF = old form, SMD = standardized mean difference.

Table 3*Average SEE and Bias for Test X for the Small, Moderate, and Large Ability Difference Conditions*

	Small ability difference (SMD = 0.03)				Moderate ability difference (SMD = 0.19)				Large ability difference (SMD = 0.28)			
Average SEE CH-linear	0.182	0.261	0.329	0.416	0.212	0.267	0.323	0.439	0.210	0.286	0.343	0.422
Average SEE Tucker	0.181	0.250	0.308	0.372	0.206	0.253	0.293	0.381	0.202	0.275	0.306	0.381
Average SEE Levine	0.190	0.292	0.389	0.576	0.225	0.302	0.403	0.618	0.226	0.318	0.435	0.590
Bias CH-linear	0.205	0.271	0.209	0.128	0.077	0.019	0.152	0.378	0.123	0.054	0.283	0.651
Bias Tucker	0.226	0.295	0.267	0.227	0.171	0.197	0.439	0.794	0.298	0.380	0.760	1.291
Bias Levine	0.184	0.249	0.144	0.085	0.067	0.237	0.200	0.325	0.073	0.340	0.327	0.333
RMSE CH-linear	0.275	0.376	0.389	0.436	0.226	0.268	0.357	0.579	0.244	0.292	0.444	0.776
RMSE Tucker	0.290	0.387	0.408	0.437	0.268	0.321	0.528	0.881	0.360	0.469	0.819	1.346
RMSE Levine	0.265	0.384	0.415	0.583	0.235	0.384	0.450	0.699	0.238	0.466	0.544	0.678

Note. The smallest numbers are shaded in gray. CH = chained, RMSE = root mean squared error, SEE = standard error of equating, SMD = standardized mean difference.

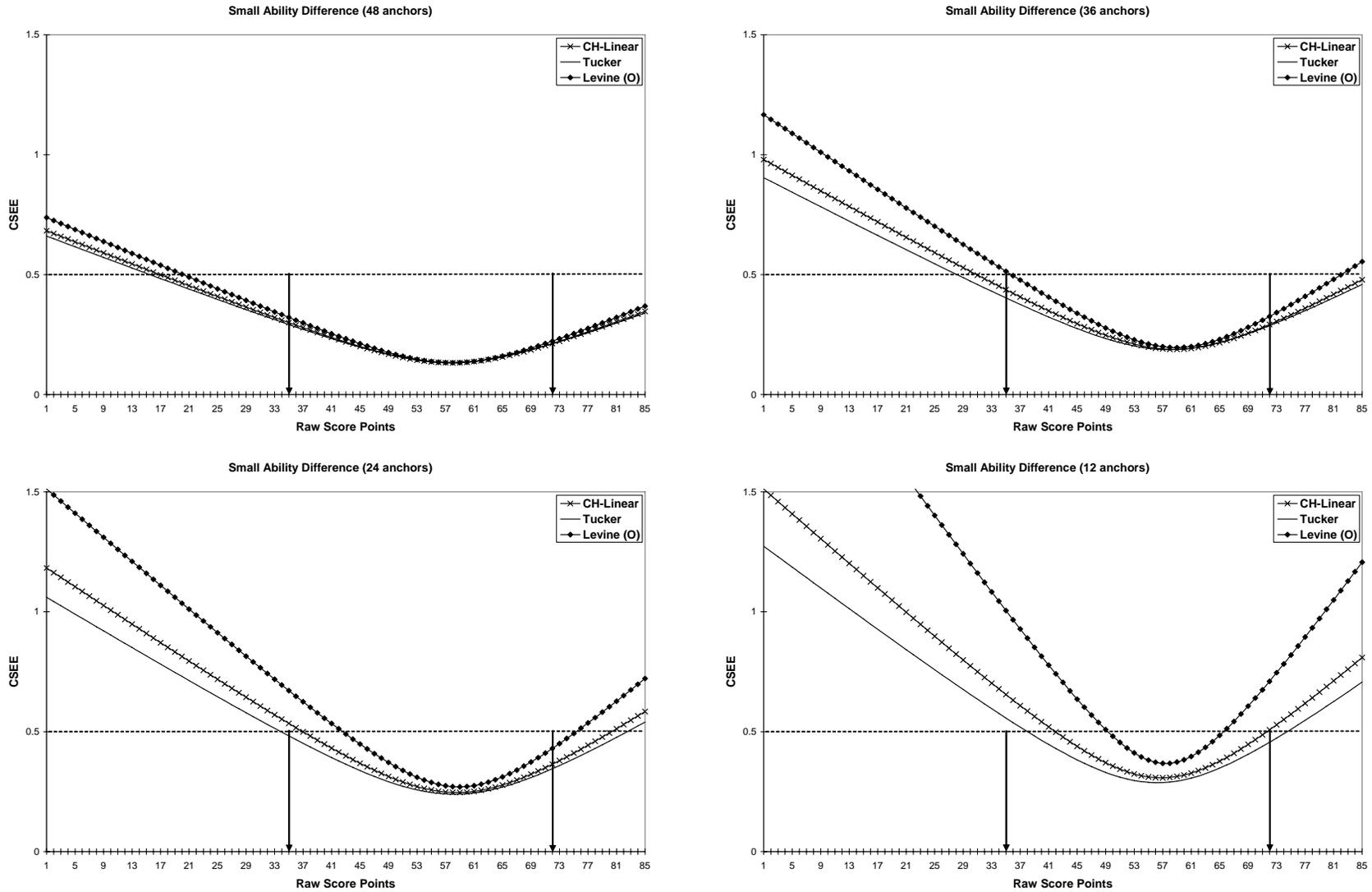


Figure 2. Conditional standard error of equating (CSEE) for small ability difference condition (Test X).

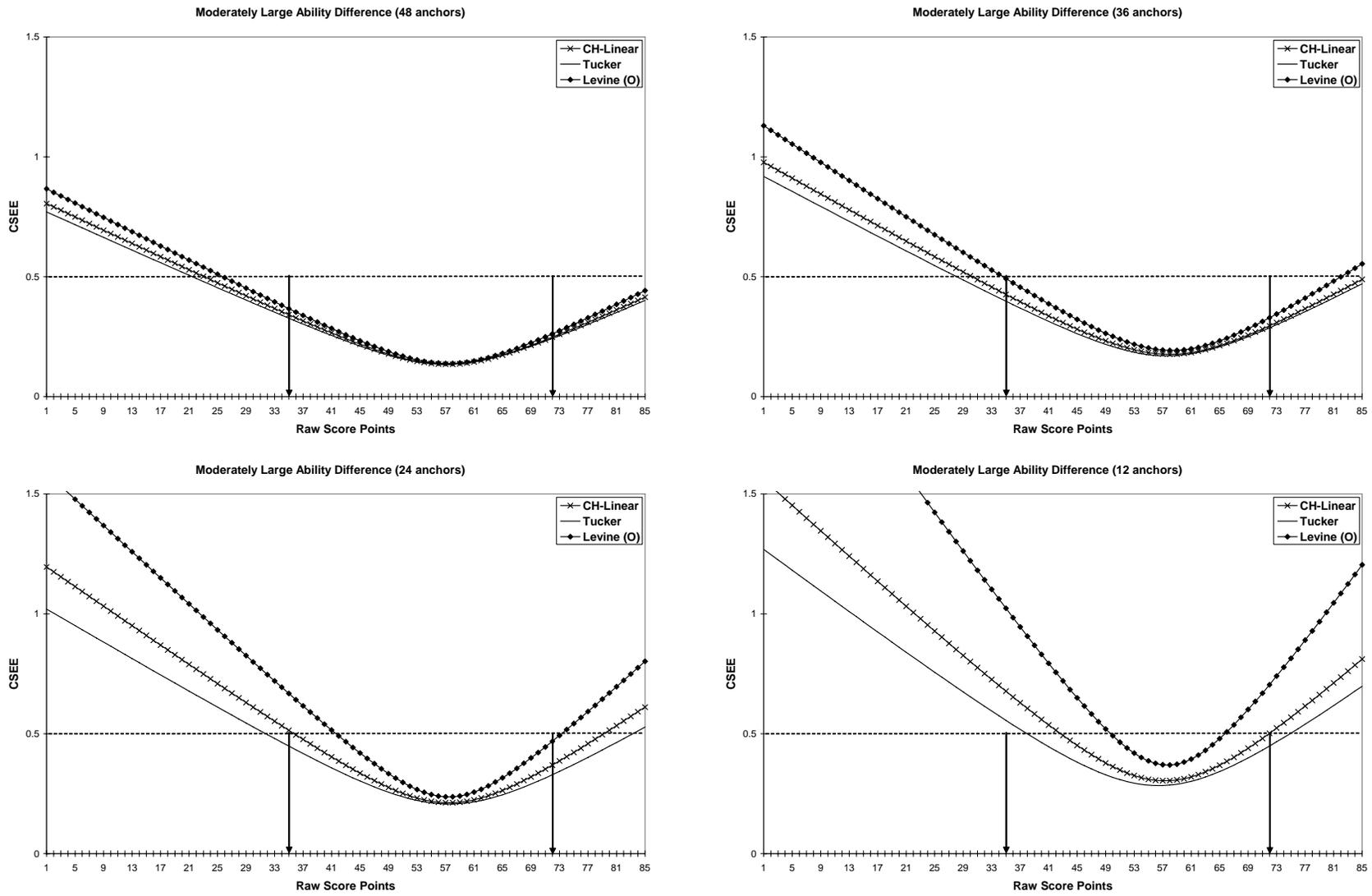


Figure 3. Conditional standard error of equating (CSEE) for moderate ability difference condition (Test X).

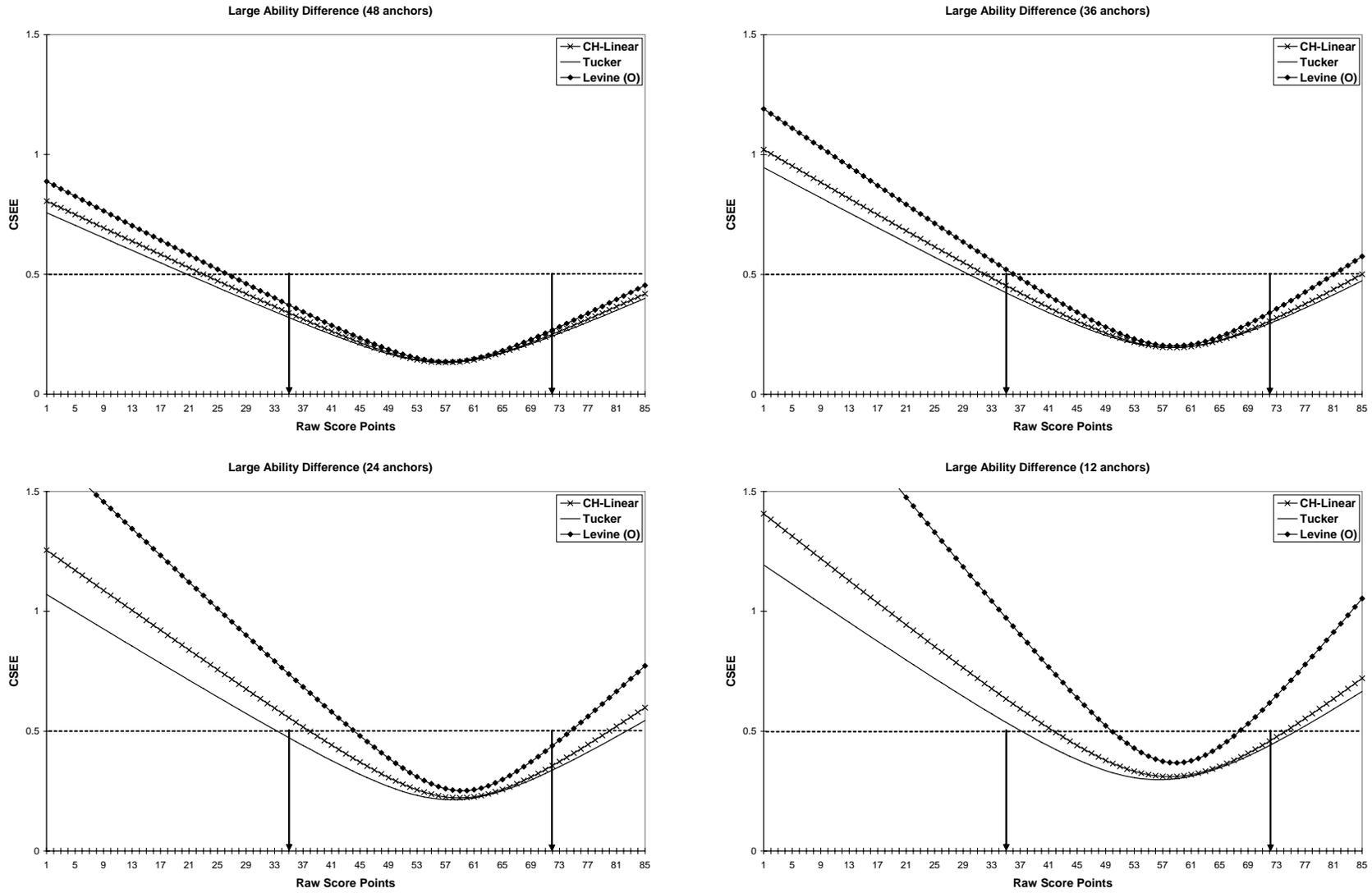


Figure 4. Conditional standard error of equating (CSEE) for large ability difference condition (Test X).

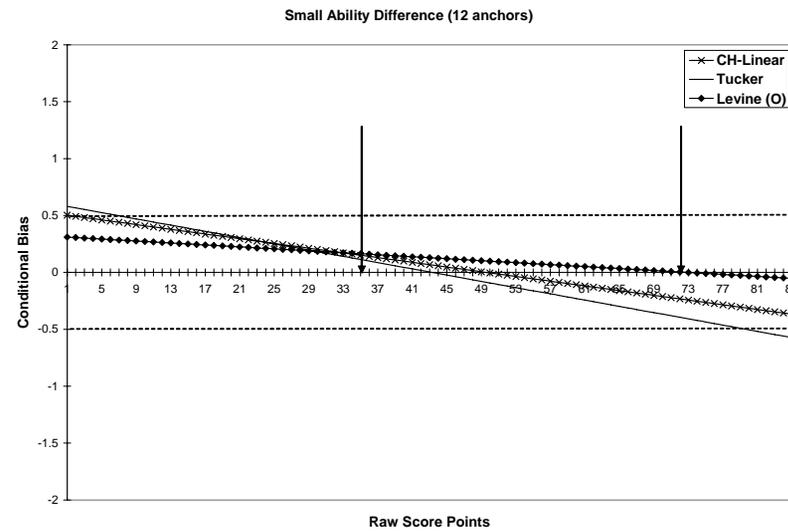
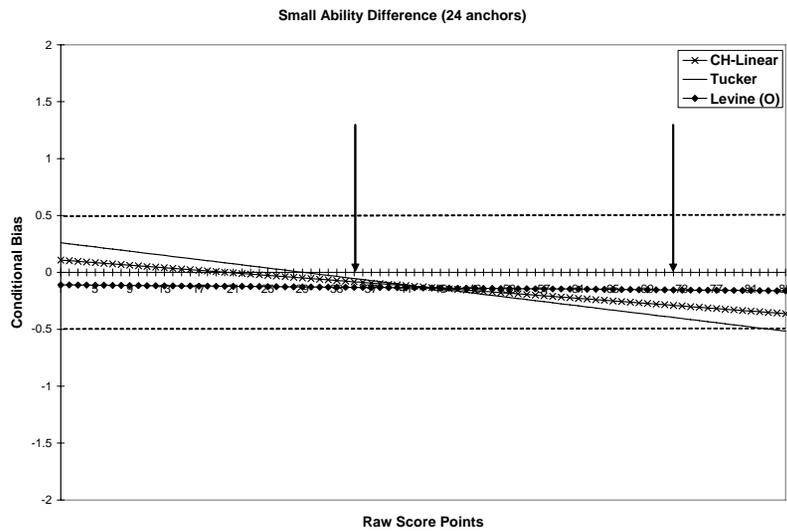
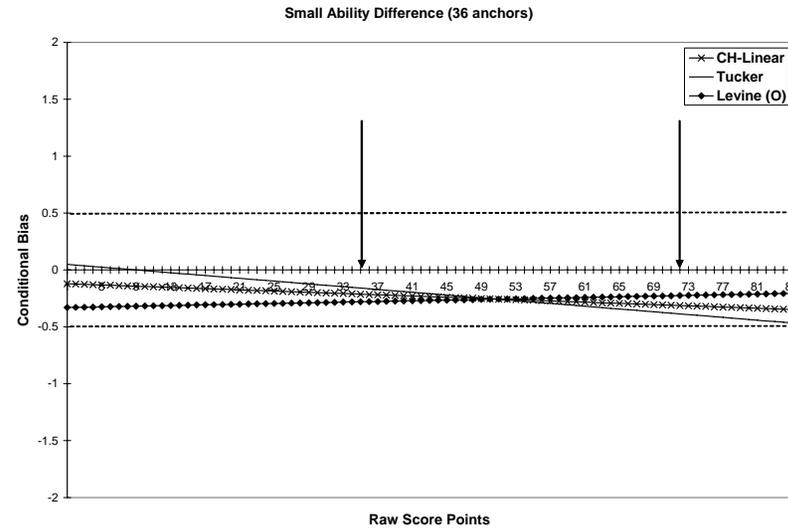
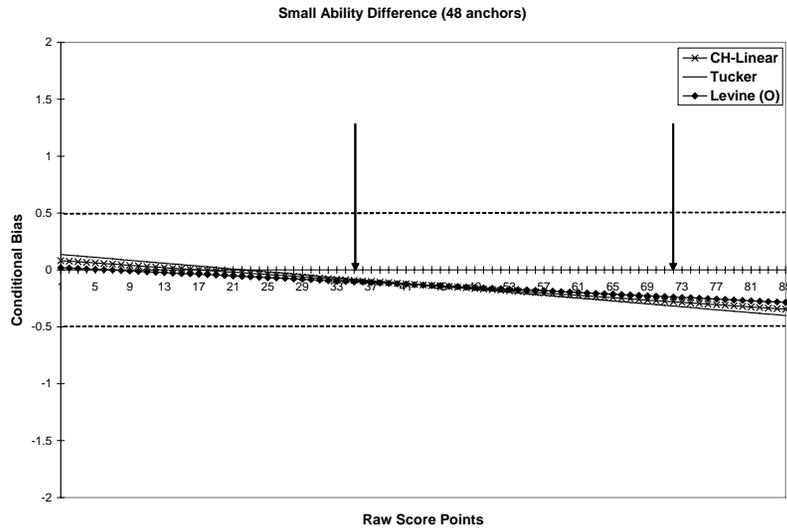


Figure 5. Conditional bias (CBias) for small ability difference condition (Test X).

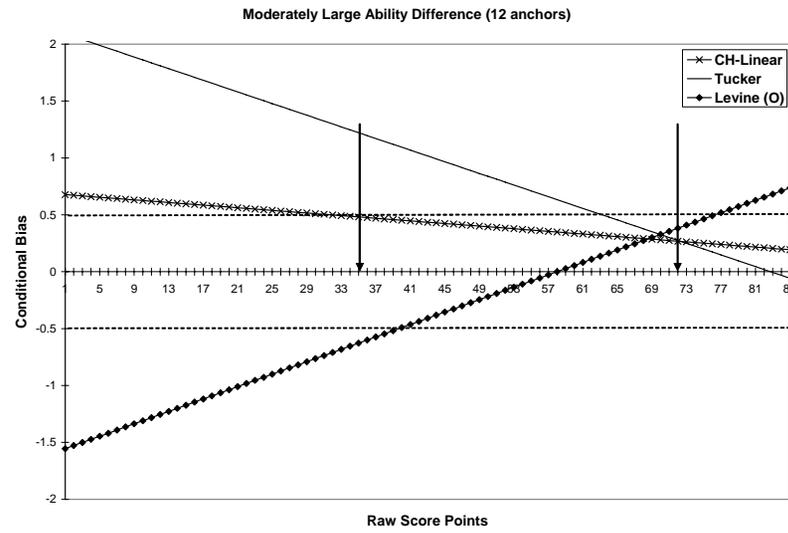
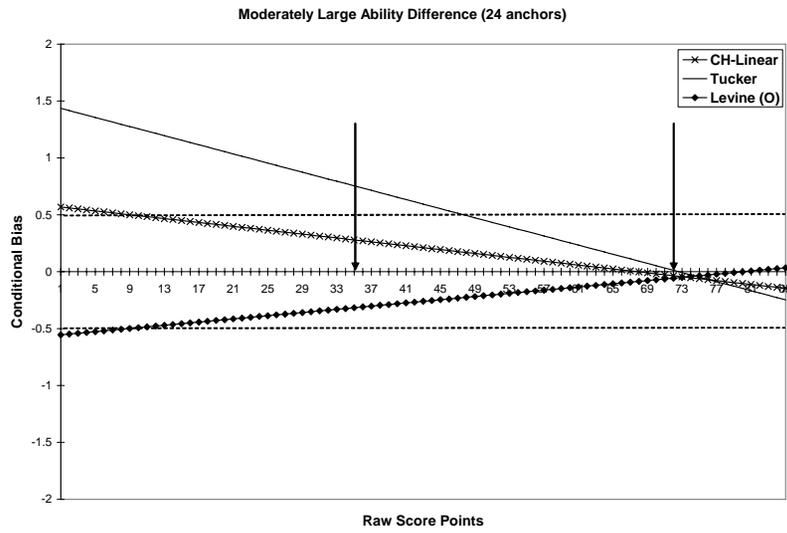
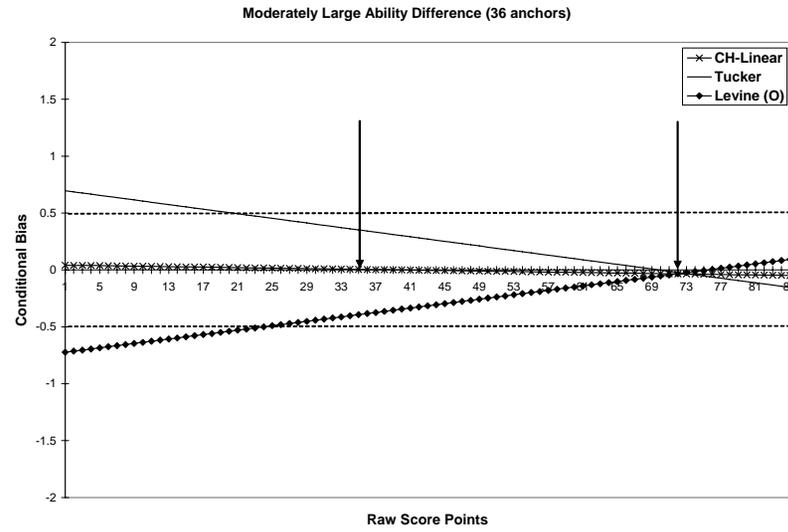
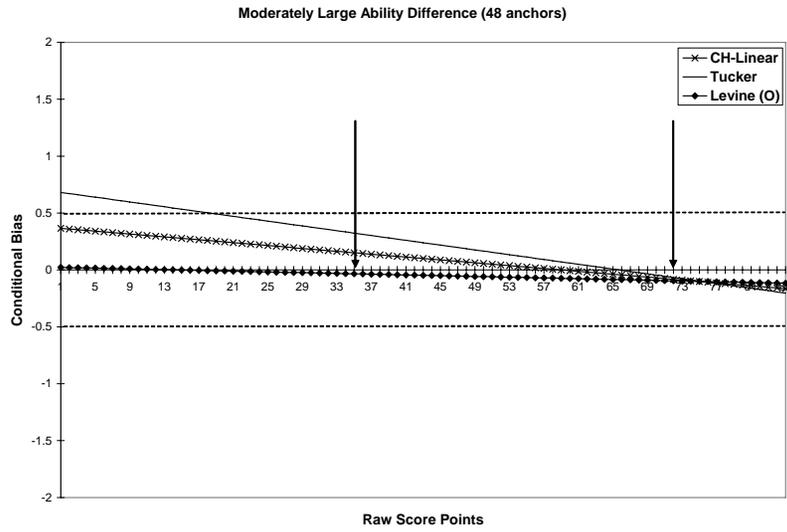


Figure 6. Conditional bias (CBias) for moderately large ability difference condition (Test X).

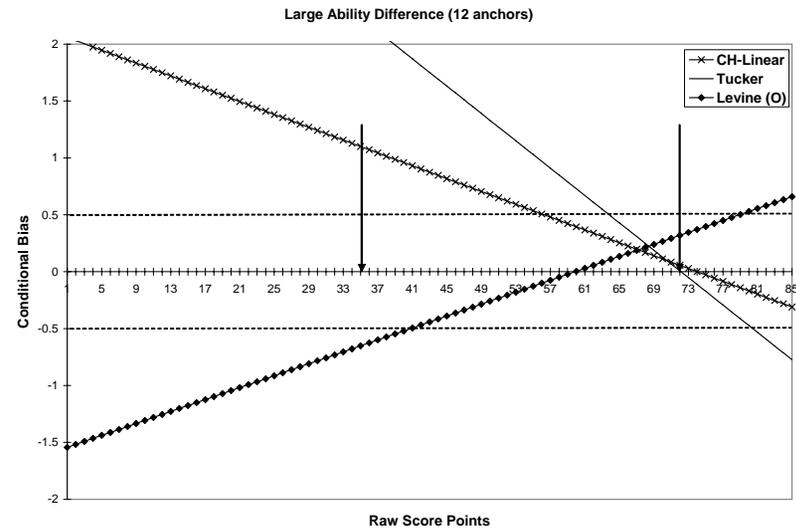
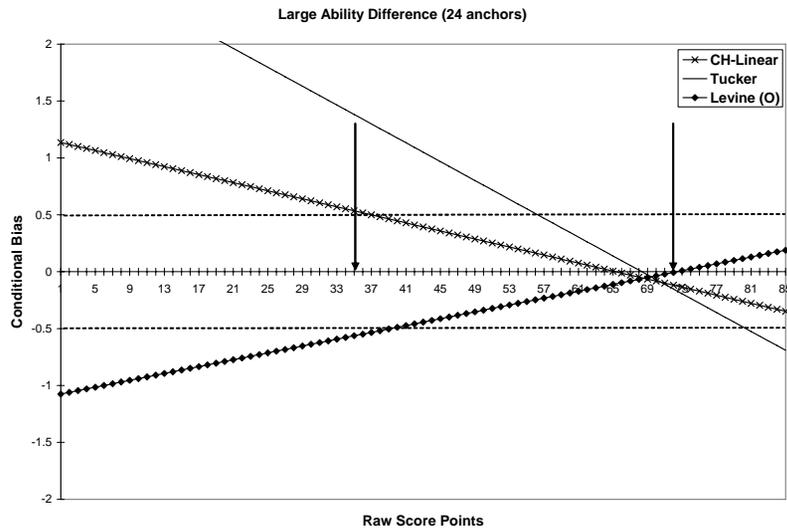
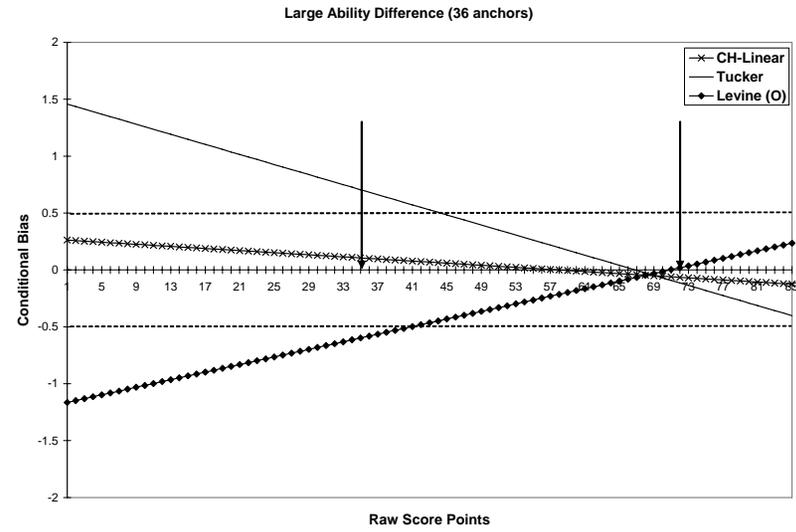
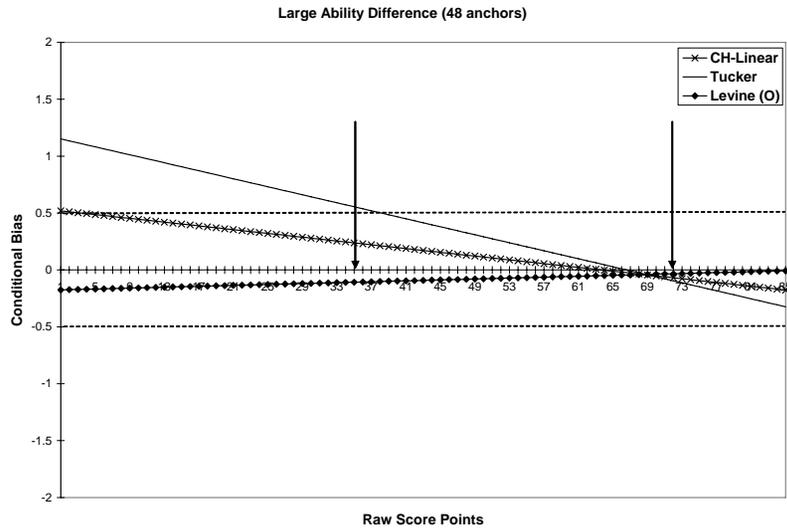


Figure 7. Conditional standard error of equating (CSEE) for large ability difference condition (Test X).

As seen in Figure 5, the conditional bias values for the three equating methods were very similar to each other in the small ability difference condition. Although the conditional bias values for the three methods start to drift apart as the anchor test length becomes smaller, the difference between them is still small enough to ignore (i.e., the conditional bias values for the three methods are all smaller than the DTM in the 35-to-72 score range). The difference in the conditional bias values of the three equating methods begins to become more apparent as the ability difference between the new- and old-form samples increases (Figures 6 and 7). As seen in Figure 6, for the moderate ability difference condition, the difference between the conditional bias values between the three equating methods increases as the length of the anchor test becomes smaller, with the Tucker method showing the largest bias. Although the conditional bias values of the Tucker method were small enough to ignore in the 48- and 36-item anchor conditions, they were not small enough to ignore for the 24- and 12-item anchor conditions for several score points in the 35-to-72 region. Finally, as seen in Figure 7, the bias in the three equating methods followed the same trend that was observed in the moderate ability difference condition, except that the conditional bias values were even further apart from each other. Also as seen in Figure 7, the chained linear method became increasingly more biased, especially for the 12-item anchor conditions. Overall, it seems that the chained linear method produces the smallest conditional bias values across most of the studied conditions.

Results for Test Y

Average SEE, bias, and RMSD results. As seen in Table 4, the average SEE for the Tucker method was always smaller than the average SEE for the chained linear or Levine methods, and the average SEE for the chained linear method was always smaller than the average SEE for the Levine method for all of the studied conditions (i.e., 12- and 20-item anchor lengths within the moderate and moderately large ability conditions). The bias for the Levine equating method was also smaller than for the Tucker and chained linear methods for all anchor length and ability conditions. Also, for all of the studied conditions, the bias for the chained linear method was smaller than for the Tucker method. In terms of RMSE, the Levine method performed slightly better than the chained linear and Tucker methods in the moderately large ability difference condition, while the chained linear method performed slightly better than the Tucker and Levine methods in the moderate ability difference condition. Finally, as seen in Table 4, the average SEEs, bias, and RMSEs for the three equating methods were larger for the

12-item anchor condition compared to the 20-item anchor condition, and this was true for both the moderate and moderately large ability difference conditions.

Table 4

Average Standard Error of Equating (SEE) and Bias for Test Y for the Moderate and Moderately Large Ability Difference Conditions

	Moderate ability difference (SMD = 0.12)		Moderately large ability difference (SMD = 0.20)	
	20	12	20	12
Number of anchor items	20	12	20	12
Average SEE CH-linear	0.205	0.276	0.194	0.270
Average SEE Tucker	0.200	0.253	0.188	0.256
Average SEE Levine	0.229	0.354	0.220	0.342
Bias CH-linear	0.082	0.195	0.180	0.418
Bias Tucker	0.174	0.345	0.317	0.629
Bias Levine	0.021	0.032	0.081	0.217
RMSE CH-linear	0.221	0.338	0.265	0.497
RMSE Tucker	0.265	0.428	0.369	0.679
RMSE Levine	0.230	0.356	0.234	0.405

Note. The smallest numbers are shaded in gray. CH = chained, RMSE = root mean squared error, SMD = standardized mean difference.

Conditional standard errors and bias. For evaluating the CSEEs and conditional bias values, it was decided to focus on the score points between the 5th and 95th percentiles. For the new form, this score roughly ranged between score points 14 and 37 in the full sample and was used to evaluate the CSEEs and conditional bias values for Test Y. As seen in Figure 8, the CSEEs for all the three equating methods were lower than the DTM for the 14-to-37 score range for the moderate ability difference condition. This was true for both the 20- and 12-item anchor length conditions. Similar to Test X results, the CSEEs for the Tucker method were slightly smaller than the CSEEs of the chained linear or Levine methods, and the CSEEs of the chained method were slightly smaller than the CSEEs of the Levine method. The CSEEs for the three equating methods also decreased as the number of anchor items increased from 12 to 20 items. As seen in Figure 8, the conditional bias values for the three equating methods were lower than the DTM for all score points for the moderate ability difference and the 20-item anchor length condition. However, for the moderate ability difference and the 12-item anchor length condition, the conditional bias values for the Tucker method were slightly larger than the DTM for some higher score points.

As seen in Figure 9, the CSEEs for all the three equating methods were lower than the DTM for the 14-to-37 score range for the moderately large ability difference condition. This was true for both the 20- and 12-item anchor length conditions. As observed before, the CSEEs for the Tucker method were slightly smaller than the CSEEs for the chained linear or Levine methods, and the CSEEs of the chained method were slightly smaller than the CSEEs of the Levine method. The CSEEs for the three equating methods also decreased as the number of anchor items increased from 12 to 20 items. As seen in Figure 9, the conditional bias values for the three equating methods were lower than the DTM for all score points for the moderately large ability difference and the 20-item anchor length condition. However, for the moderately large ability difference and the 12-item anchor length condition, the conditional bias values for the Tucker method were larger than the DTM for all score points. The conditional bias for the chained linear method were mostly under the DTM in the 14-to-37 score region, except for a few score points in the lower score region, and the conditional bias for the Levine method was under the DTM for all score points in the 14-to-37 score region.

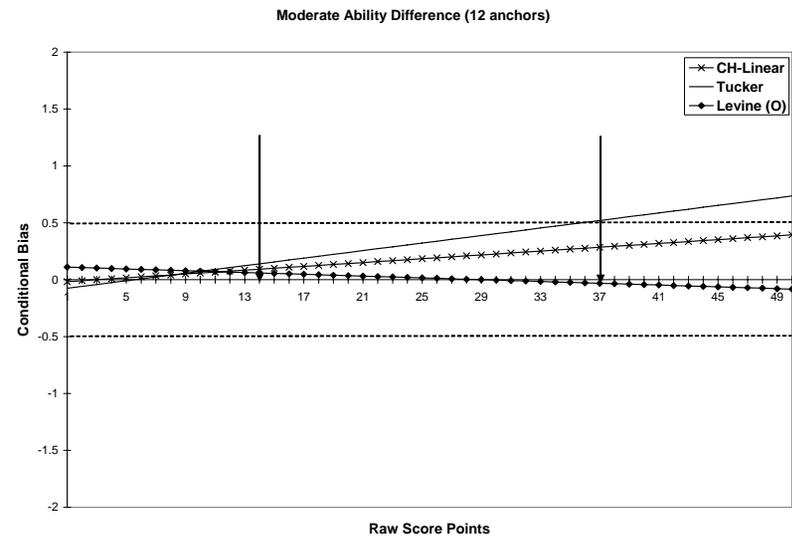
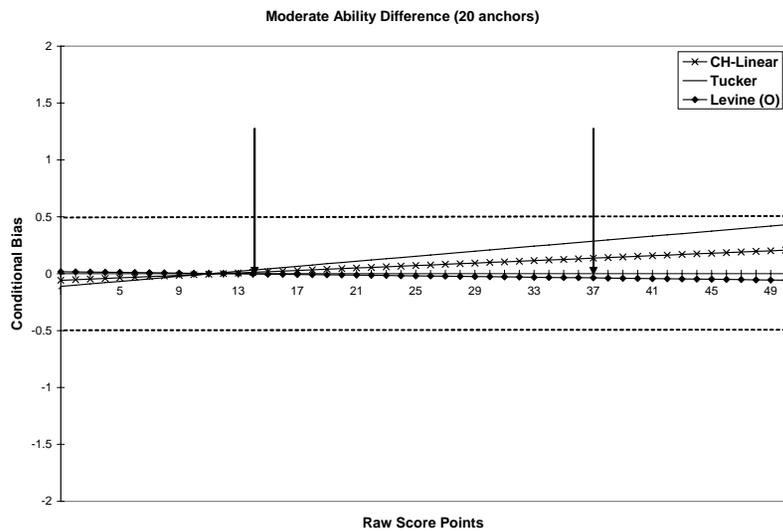
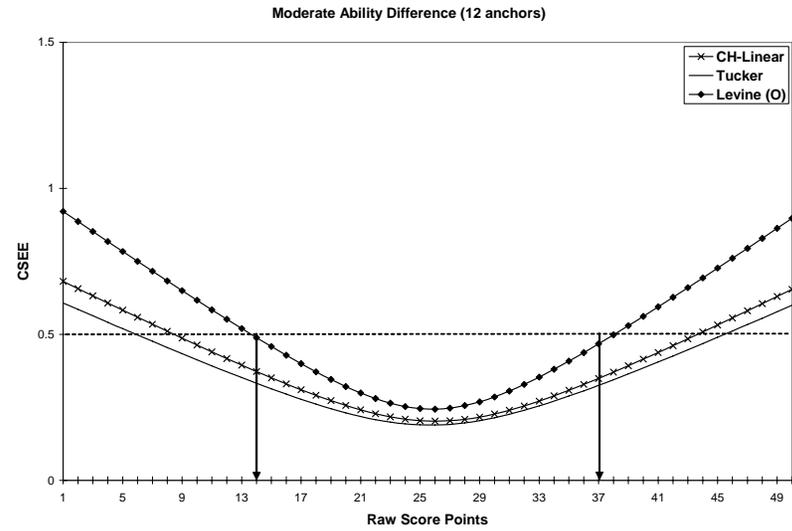
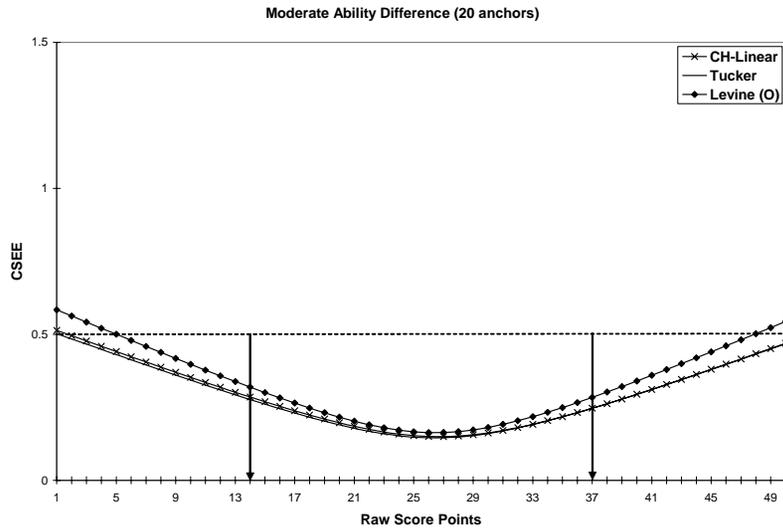


Figure 8. Conditional standard error of equating (CSEE) and conditional bias (CBias) for moderate ability difference condition (Test Y).

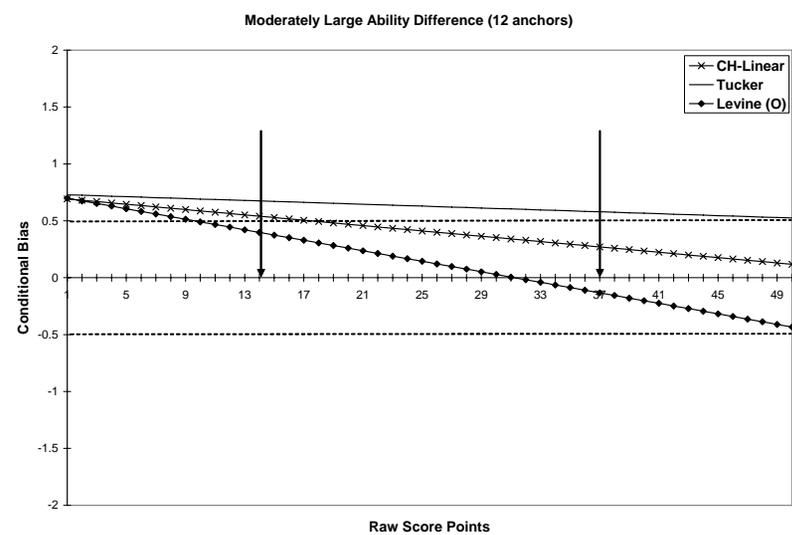
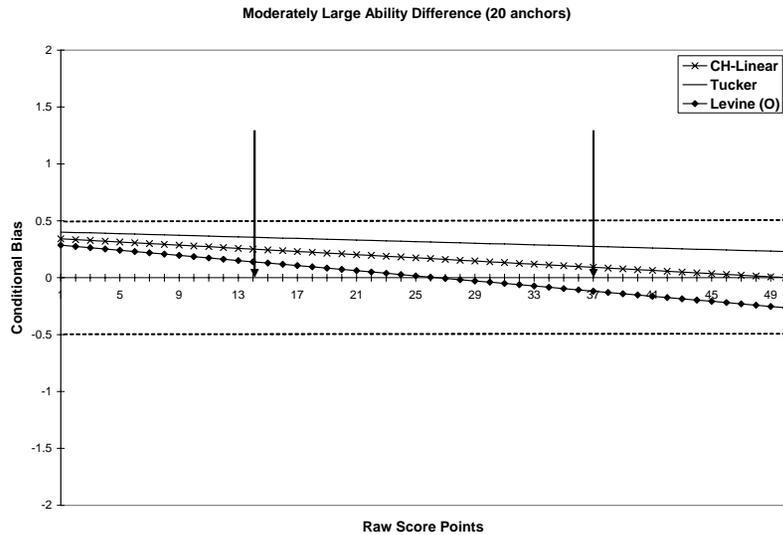
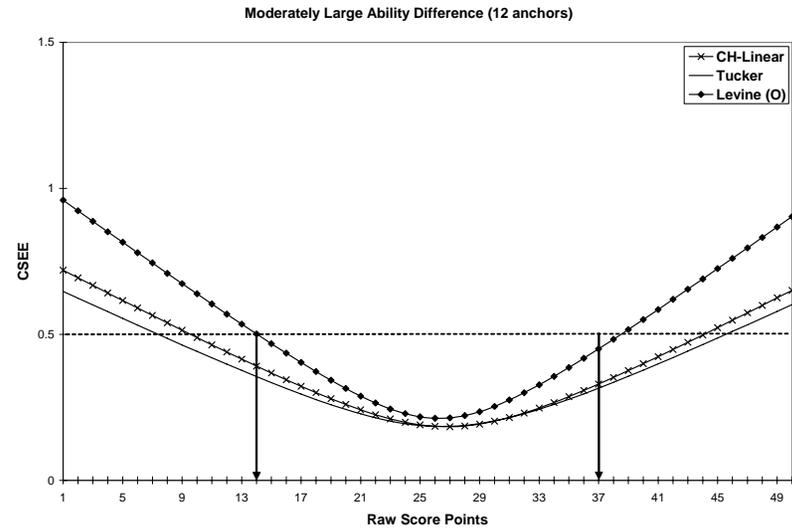
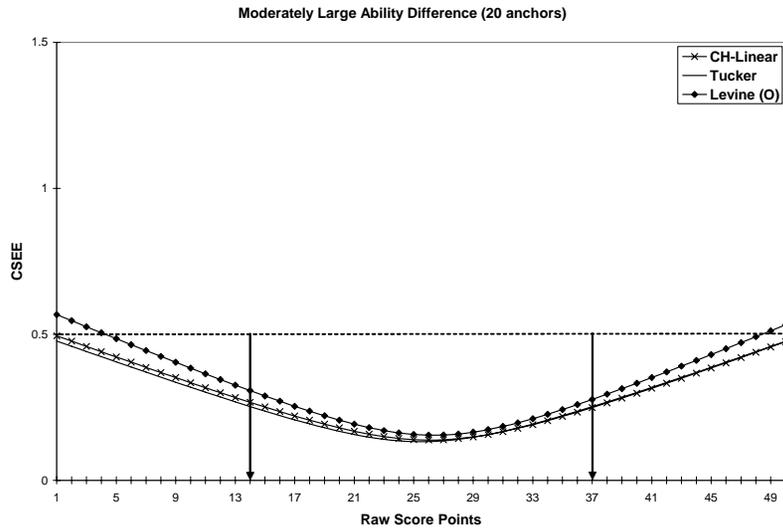


Figure 9. Conditional standard error of equating (CSEE) and conditional bias (CBias) for moderately large ability difference condition (Test Y).

Discussion and Conclusion

Previous studies based on real data have found that chained and PSE equating methods produce somewhat different results when the new- and old-form samples differ in ability. Since testing programs often administer tests at different time points, known as administrations, and test takers from these different administrations often differ in ability, this study examined the effect of small to large ability differences on the results of chained linear, Tucker, and Levine equatings. The effect of anchor length was also varied (from large to small) to examine its effect on the equatings.

The results for Test X showed that the average and conditional random equating error was always lower for the Tucker method across all of the studied conditions. A similar finding was reported in Kolen and Brennan (2004), where the Tucker method produced the smallest random equating error when compared to the Levine or frequency estimation methods. However, in terms of equating bias and overall equating error (i.e., the RMSE), the chained linear method produced good results (i.e., lower bias and RMSE) for a majority of the conditions in the moderate and large ability difference conditions. Under the moderate and large ability difference conditions, Levine equating also produced good results for some cases. The Levine method also produced good results in terms of bias for the small ability difference condition, and chained linear produced lower overall error in this condition. The Tucker method performed the worst in terms of bias and RMSE for all conditions. This supports the findings from previous research that showed that PSE equatings were more biased when new- and old-form samples differed in ability and the anchor-to-total correlation was not perfect (Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2006). Finally, the number of anchor items affected random equating error, which increased as the number of items in the anchor decreased. The bias values followed a less predictable pattern across the different anchor length conditions, especially for chained linear and Levine equating.

The results of Test Y showed that the average and conditional random equating error was always lower for the Tucker method across all of the studied conditions. However, in term of equating bias and RMSE, the chained linear and Levine methods produced lower bias and RMSE as compared to the Tucker method. The Levine method had the lowest bias across all conditions, while in terms of RMSE, both the chained and the Levine methods did equally well. As before, the Tucker method produced the largest bias and RMSE compared to the other methods. Finally,

the number of anchor items affected random equating error, which increased as the number of items in the anchor decreased. The bias and RMSE also tended to be generally larger when the number of anchor items was small.

Based on these findings, either the chained linear, Tucker, or Levine equatings may be used when the difference in the new- and old-form samples is small and the correlation between the anchor and total test is at least moderately high (e.g., high 70s and above). However, when the new- and old-form samples differ in ability, chained linear equating seems preferable because this method produced the lowest RMSE in a majority of the studied conditions (i.e., 9 out of 12 conditions for Test X, and 2 out of 4 conditions for Test Y). Levine also produced good results, especially for Test Y, and may be a viable option when samples differ in ability. Finally, based on these results, it seems reasonable not to use the Tucker equating method when new- and old-form samples differ in ability, unless the correlation between the anchor and the total test is very high (e.g., 0.90 and higher). Under these high anchor-to-total correlation conditions, chained linear and Levine equating also produce good equating results and would be reasonable choices.

Why Did Tucker Equating Perform Poorly? A Partial Explanation

How well an equating method performs partially depends on how well the assumptions specific to a particular equating method are met. In this study, it is reasonable to assume that in the conditions where the Tucker method performed poorly compared to the other methods, the assumptions made by the Tucker method were likely violated. What are these assumptions? For the Tucker method, the assumption is that the *linear regression* of the total on the anchor score in the new form is the same across the new- and old-form samples, and the *linear regression* of the total on the anchor score in the old form is the same across the new- and old-form samples. The assumptions for chained linear and Levine are similar, except that the *linear regression* term in Tucker equating is replaced by the *observed scaling* in chained linear equating, and *true score regression* in Levine equating (see von Davier, 2008, for a detailed discussion of these assumptions and their derivations). These assumptions for Tucker, chained linear, and Levine equating can be written as shown in Equations 1, 2, and 3, respectively. In these formulas, x and A represent the total and anchor scores.

$$\text{Regression}_X(A) = \mu_X + \text{corr}_{XA} \frac{\sigma_X}{\sigma_A} (A - \mu_A) \quad (1)$$

$$\text{Observed Scaling}_X(A) = \mu_X + \frac{\sigma_X}{\sigma_A} (A - \mu_A) \quad (2)$$

$$\text{True Score Regression}_X(A) = \mu_X + \frac{\sqrt{\text{rel}_X} \sigma_X}{\sqrt{\text{rel}_A} \sigma_A} (A - \mu_A) \quad (3)$$

In actual test administrations, the new form is typically taken by the new-form sample, and the old form is taken by the old-form sample. So how would it be possible to obtain an anchor to total relationship for the new form using the old-form sample, or the same relationship for the old form using the new-form sample? In reality, it would be difficult to examine this, but the current study (especially for Test X) was designed in a way that makes such an examination possible. For the purpose of illustration, the 24-anchor moderately large ability condition for Test X was used. Recall that for Test X, Forms X1 and X2 were created from a single larger form (Form X). After the two forms were created, two groups of examinees, which differed moderately in ability, were each assigned to either the new or the old form. But since the two forms were created from Form X and all examinees took Form X, it essentially means that even though the equatings were conducted by assigning a specific group to either Form X1 or Form X2, all examinees responded to Form X1 and Form X2. This enabled us to test the assumptions of the three methods exactly. For example, to test the Tucker equating assumption, the anchor-to-total regression for Form X1 was calculated using the new sample. Since the sample that took Form X2 also had responses for Form X1, the anchor-to-total regression for Form X1 was also calculated using the old-form sample. Similarly, the anchor-to-total regression for Form X2 was calculated using the old-form sample and, since the sample that took Form X1 also had responses for Form X2, the anchor-to-total regression for Form X2 was also calculated using the new-form sample. The assumption of Tucker equating can be said to be adequately met if the difference between these two regressions across the new- and old-form samples is small. This procedure is repeated to evaluate the assumptions of chained linear and Levine equating.

In the appendix, Figure A1 shows the difference in the anchor-to-total relationship in Form X1 using the new- and the old-form samples, and Figure A2 shows the difference in the

anchor-to-total relationship in Form X2 using the new- and the old-form samples. A small difference would indicate that the assumptions of a particular equating method were adequately met. As seen in Figure A1, this difference was smallest for chained linear method and somewhat larger for the Tucker and Levine methods. As seen in Figure A2, this difference was smallest for the Levine method and somewhat larger for the Tucker and chained linear methods. Based on this information, one would expect the Tucker method to perform the worst in terms of equating bias, and the chained linear and Levine methods to produce some bias but in general produce lower bias than the Tucker method. As observed in Figure 6, this expectation seems accurate. Tucker performed the worst in terms of bias, and chained linear and Levine had similar amounts of bias, although in opposite directions. These results suggest that although the Tucker method performed poorly in this condition, this poor performance is not attributable to any inherent pervasive flaw in the Tucker method. It simply means that the assumptions of the Tucker method were not adequately met.

Can Tucker Equating Perform Better Than Chained Equating in Some Conditions?

Following the concluding statement in the previous section, it is reasonable to expect Tucker equating to produce good results if the assumptions of the method were adequately met. According to Livingston (2004), the Tucker equating method uses the anchor score only to the extent it correlates with the total score, and when the correlation between the anchor and total scores is not perfect, the Tucker method assumes that the new- and old-form samples are more similar in ability than they may actually be. This may introduce some bias when the actual samples are different in ability, thereby resulting in an equating where the adjustment is smaller than actually needed. However, this same feature may sometimes lead Tucker to perform better than other methods, such as chained linear equating. Recall, that the chained linear equating method involves a scaling of the total-to-anchor scores in the new form and the old form and then chaining these scores together. However, this anchor-to-total scaling implicitly assumes that the correlation between the anchor and total scores is perfect. Therefore, in situations where the anchor score is weakly correlated to the total score, chained equating may lead to a less accurate equating than the Tucker method (although note that no equating will produce good results when the correlation between the anchor and the total scores is very low). For the purpose of illustration only, scores on a language measure were equated using a social studies external anchor. Similar to Test Y, the language test was equated to itself using two different samples as

new- and old-form samples, and the identity equating was used as a criterion. The new- and old-form samples were chosen to be of similar ability ($SMD = 0.04$). The correlation between the language and social studies scores was 0.5. As seen in Figure A3, the chained linear method resulted in a slightly larger bias (also larger than the DTM) than the Tucker method in the lower score region. This bias may even be larger when the correlation between the anchor and total scores is lower than 0.5, as was observed for this data. The overall bias and RMSE for the Tucker method were 0.085 and 0.206, respectively, and for the chained linear method were 0.095 and 0.236, respectively, indicating that at the total score level, Tucker equating produced slightly more accurate results than chained linear equating.

One possible reason for this result is that since the correlation between the anchor and total test was quite low, the Tucker equating adjusted for form difficulty as if the new- and old-form samples were more similar in ability than was indicated by the anchor test (see Livingston, 2004). In this case, since the new- and old-form samples were already very similar, assuming that the two samples were similar in ability did not introduce any additional bias. Furthermore, by using less of the anchor, which is weakly correlated to the total test, the Tucker method provides a slightly better equating than the chained linear method, which uses all the information provided by the anchor to make the adjustment. This result may be informative for testing programs which methods adjust scores on the basis of anchors that may not be well correlated with the total scores. For example, constructed-response tests consisting of very few items are often equated through an external anchor test where the correlation between the anchor and total scores is quite low (see Puhan, von Davier, & Gupta, 2008, for an example).

Limitations and Future Research

In this study, the correlation between the anchor test and total test was varied to examine its effect on CE and PSE. However, even the smallest anchor test resulted in a moderate anchor-to-total correlation (i.e., in the high 70s for Test X and low 80s for Test Y). It was not possible to further reduce the number of items because that would have resulted in an anchor test that did not adequately represent the total test in terms of content. So the effect of very low anchor-to-total correlations on the CE and PSE results could not be examined. One way to create an anchor test with a low anchor-to-total correlation and at the same time not significantly reduce the number of items in the anchor test is to build an anchor test using items that would represent the content of the total test but have low biserial correlations (i.e., correlation between the item and

the total test score). Future studies can use this approach to create anchor tests that do not correlate well with the total test, and examine their effects on CE and PSE.

Finally, this study used data from only two tests. Although it seems reasonable to expect similar results if other tests were used, it is possible, under some unique circumstance, for the results to be different from what was observed in this study. Therefore, more studies of a similar nature need to be conducted and if those studies also find similar results, then the finding that CE tends to produce more accurate equating results than PSE when new- and old-form samples differ in ability and when anchor-to-total correlations are weak can be more strongly claimed. In the meantime, results of this study can serve as a guide to those who must equate when anchor-to-total correlations are low and new- and old-form samples differ in ability.

References

- Dorans, N. J. (Ed.). (1990). Selecting samples for equating: To match or not to match [Special Issue]. *Applied Measurement in Education*, 3(1).
- Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT® and PSAT/NMSQT®* (ETS Research Memorandum No. RR-94-10). Princeton, NJ: ETS.
- Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement*, 50, 61–71.
- Kim, S., Livingston, S. A., & Lewis, C. (2009). *Investigating the effectiveness of collateral information on small-sample equating* (ETS Research Rep. No. RR-09-14). ETS: Princeton, NJ.
- Kim, S., von Davier, A., & Haberman, S. (2006). *An alternative to equating with small samples in the non-equivalent groups anchor test design* (ETS Research Rep. No. RR-06-27). ETS, Princeton, NJ.
- 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. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement*, 30(1), 23-29.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.
- Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3(1), 73–95.
- Marco, G. L, Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp. 147–176). New York, NY: Academic.
- Puhan, G., von Davier, A. A., & Gupta, S. (2008). *Impossible scores resulting in zero frequencies in the anchor test: Impact on smoothing and equating* (ETS Research Rep. No. RR-08-10). ETS. Princeton, NJ.
- 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.
- von Davier, A. A. (2008). New results on the linear equating methods for the nonequivalent-groups design. *Journal of Educational and Behavioral Statistics*, 33(2), 186–203.

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

Appendix

Difference in the Anchor-to-Total relationship in Forms X1 and X2

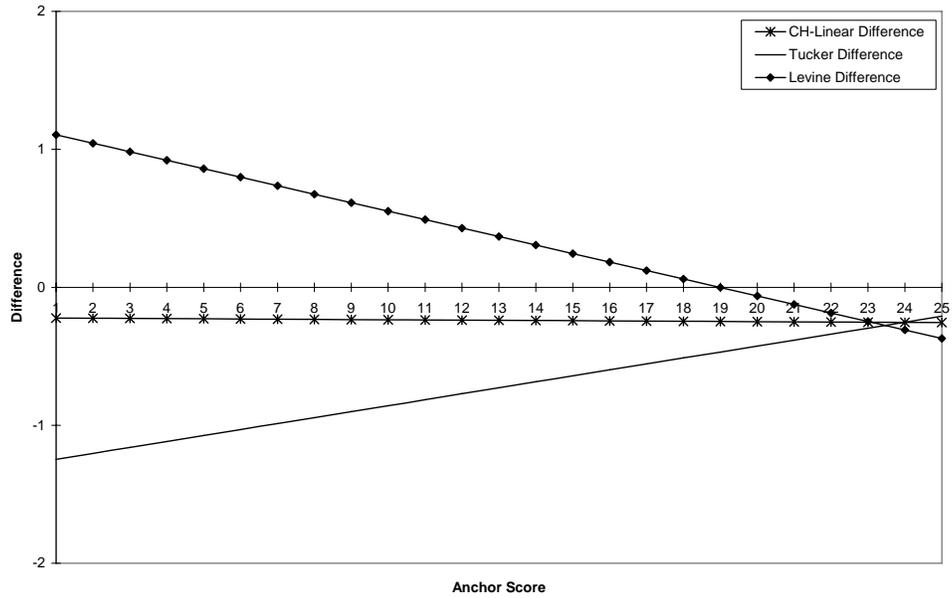


Figure A1. Difference in new-form total-to-anchor relationship obtained using new- and old-form samples.

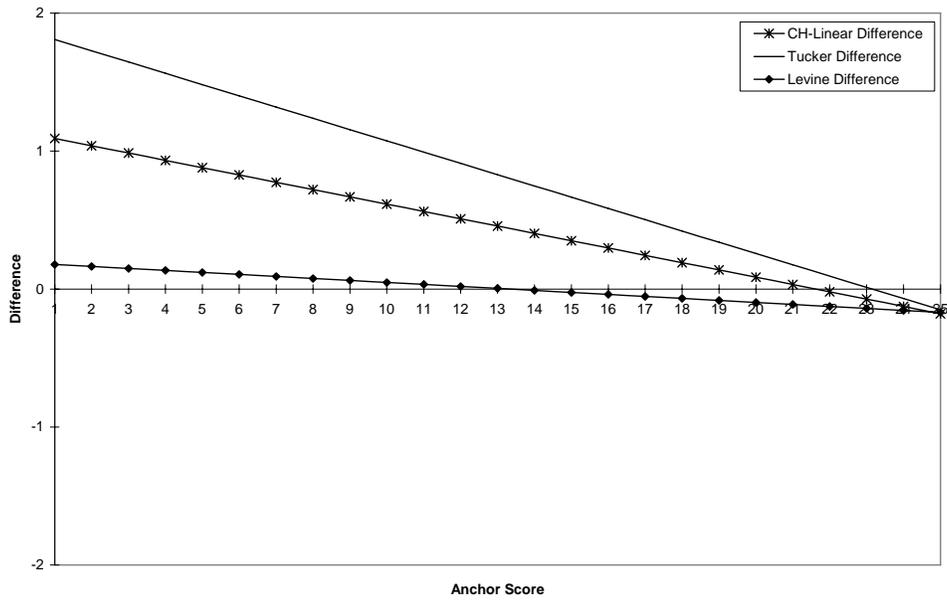


Figure A2. Difference in old-form total-to-anchor relationship obtained using new- and old-form samples.

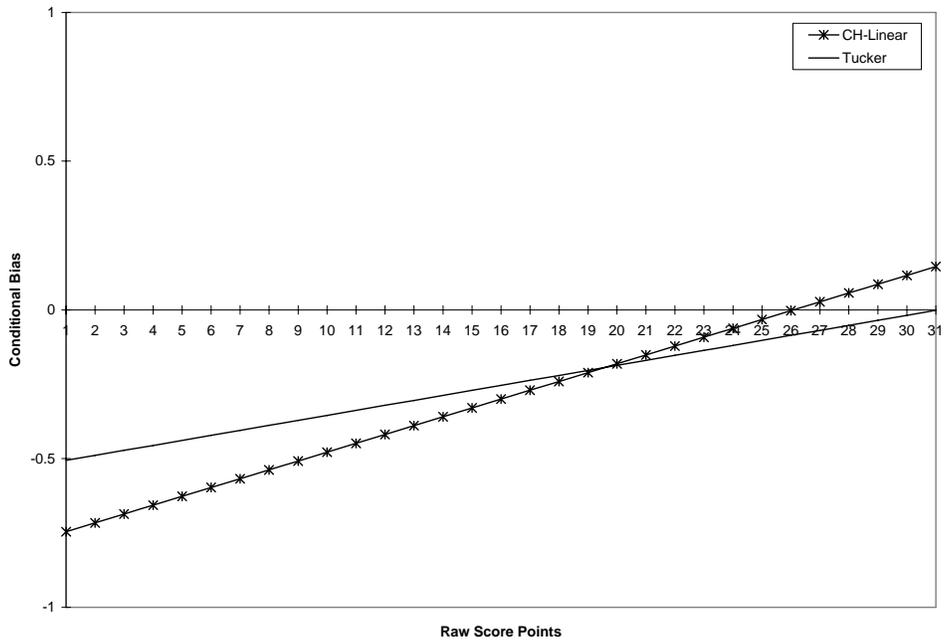


Figure A3. An example in which Tucker showed less bias than was shown by chained linear equating.