Methods of Linking With Small Samples in a Common-Item Design: An Empirical Comparison

Sooyeon Kim and Samuel A. Livingston

December 2009

ETS RR-09-38

Methods of Linking With Small Samples in a Common-Item Design:
An Empirical Comparison

Sooyeon Kim and Samuel A. Livingston
ETS, Princeton, New Jersey

December 2009

Abstract

A series of resampling studies was conducted to compare the accuracy of equating in a common-item design using four different methods: chained equipercentile equating of smoothed distributions, chained linear equating, chained mean equating, and the circle-arc method. Four operational test forms, each containing more than 100 items, were used for the equating, with new-form samples of 100, 50, 25, and 10 examinees and reference-form samples three times as large. Accuracy was assessed in terms of the root-mean-squared difference (over 1,000 replications) of the sample equatings from the criterion equating. Overall, chained mean equating produced the most accurate results for low scores, but the circle-arc method produced the most accurate results, particularly in the upper half of the score distribution.

Key words: Circle-arc method, small-sample equating, chained mean method, equipercentile method, equating bias, equating error

Acknowledgments

We would like to thank Mary C. Grant, Rick Morgan, and Daniel Eignor for their many helpful comments on the earlier draft of this paper. The authors also gratefully acknowledge the editorial assistance of Kim Fryer.

Table of Contents

List of Tables

List of Figures

In testing programs, multiple forms of a single test are used in different administrations for test security reasons. However, as well specified as the test development process may be, differences in the statistical difficulties of the alternate forms often occur. An equating procedure can adjust for these differences so that scores on the test forms are comparable, but as with other statistical procedures the equating of test scores is subject to sampling effects. If the sample is large and representative, the equating relationship in the sample is likely to represent accurately the equating relationship in the population. The smaller the sample, the more likely it is that the equating function computed for that particular sample will differ substantially from that of the population.

The relationship between equating accuracy and sample size has been shown theoretically in formulas for the standard error of equating (Kolen & Brennan, 2004) and empirically in the results of resampling studies (Kim, von Davier, & Haberman, 2008; Parshall, Houghton, & Kromrey, 1995; Skaggs, 2005). Nevertheless, many practitioners must use small data sets to determine the equating transformation to be used in an operational testing situation, because many testing programs must report comparable scores for a new edition of an established test in a timely manner, regardless of sample size. In these and other cases test equating sometimes must be performed with samples as small as 20 to 30 examinees per test form.

Several studies have been conducted to explore potential methodological alternatives to equating with small samples, such as empirical Bayes (EB) estimation incorporating collateral information, the synthetic linking function incorporating the identity function, and presmoothing of score distributions (Kim et al., 2008; Livingston 1993; Livingston & Lewis, 2009). The common idea behind these methods is that any resulting increase in systematic error caused by using collateral information, averaging with the identity, or smoothing the distributions is more than offset by the decrease in random error. Empirical investigations, particularly for the synthetic method (Kim, von Davier, & Haberman, 2007, in press) and the EB method (Kim, Livingston, & Lewis, 2008, in press), have shown that the usefulness of these methods depends heavily on the sample size and the difference in the difficulty of the forms being equated. Overall, with equating samples of 30 or fewer examinees, those methods have yielded improved accuracy of equating when compared with traditional equating methods (e.g., the chained linear method).

Recently, Livingston and Kim (2008, 2009) proposed an equating procedure for estimating a curvilinear equating function from small-sample data. This method, called the circle-arc method, constrains the estimated equating curve to pass through two prespecified endpoints and a middle point determined from the data. The estimate is obtained by decomposing the equating transformation into a linear component and a curvilinear component and using a circle arc as the model for the curvilinear component. In a small-scale resampling study using one pair of test forms with samples of 25 examinees in a common-item equating design, the circle-arc method produced more accurate results than the other methods investigated, which included mean equating, three linear equating methods, and equipercentile equating of smoothed distributions. The research described in this paper attempts to determine the extent to which this result would generalize to other tests and other sample sizes.

The authors have conducted similar studies to compare methods of equating in an equivalent-groups design with sample sizes of 400, 200, 100, and 50 examinees taking each form (Livingston & Kim, in press). The methods investigated were circle-arc estimation, mean equating, linear equating, and equipercentile equating of smoothed distributions. For samples of 200 or fewer examinees, the circle-arc method produced the most accurate results, particularly above the 75th percentile of the score distribution. (Mean equating equaled it in accuracy for scores below the 25th percentile, and all methods performed about equally well between the 50th and 75th percentiles.)

The sample sizes required to ensure a given level of equating accuracy differ for various equating designs. An equivalent-groups design requires a much larger sample than an anchor-test design, which requires a larger sample than a single-group design. Perhaps the most commonly used equating design is the common-item design, in which each test form is administered to one of two groups that may differ in ability.

The purpose of the present study was to compare the accuracy of four methods for equating in a common-item design, with small equating samples drawn from populations that differ in ability. The investigation consisted of a series of resampling studies. The methods investigated were chained equipercentile equating of smoothed distributions, chained linear equating, chained mean equating, and the circle-arc method. Following Angoff's "commonly accepted definition of equivalent scores" (1971, p.563; 1984, p. 86), we considered an equating based on two samples of examinees to be accurate to the extent that it matched the equipercentile equating in the populations from which those samples were drawn.

Method

Data

This study involved the equating of four pairs of test forms, through common items. Those test forms were research forms constructed for the study. Each pair of research forms was constructed by extracting items from a single operational test form. Each of the four operational forms included at least 110 items and had been taken by at least 10,000 examinees. Table 1 presents, for each of the four operational test forms, the number of items in the test, the number of examinees taking the form at two different administrations, and the mean and standard deviation of their raw (number-correct) scores, along with the standardized mean differences between the scores at the two administrations. On each test, the examinees tested at one administration tended to differ systematically in ability from those tested at the other administration.

Table 1

Four Operational Test Forms Used As Item Pools

Operational form	Number of items	Administration 1			Administration 2			SMD
		N	*M*	*SD*	*N*	*M*	*SD*	
1	117	6,580	80.08	14.39	6,467	76.91	15.33	+ 0.21
2	118	6,180	82.33	16.05	8,818	84.42	15.02	- 0.14
3	110	5,475	84.60	10.29	4,776	82.24	10.93	+ 0.22
4	110	7,130	77.95	12.31	5,434	81.51	11.45	- 0.30

Note. SMD = Standardized mean difference, Administration 1 minus Administration 2.

Using each of the four selected operational test forms as an item pool, we created a pair of research forms from each operational form. The two research forms in each pair shared at least 23 items in common, were equal in length (around 60% as long as the parent form), and were parallel in content, but were unequal in difficulty. The correlations between the total score and anchor score ranged from .85 to .91.

Each pair of research forms was equated in two sets of resampling studies. In one set of studies, the full group of examinees tested at the first administration was designated as the new-form population; the group tested at the second administration was designated as the reference-form population. In the other set of resampling studies using the same pair of research forms, the group designations were reversed. In both sets of studies, the target population for equating was formed by combining the two administration groups.

Table 2 shows a statistical comparison of the two research forms created from each of the four operational forms. The comparison is based on data from the full population, the combined group of all examinees at the two administrations. In two of the four pairs of research forms, the new form was more difficult than the reference form; in the other two pairs, the new form was easier than the reference form. The difference between the means of the combined group on the two research forms varied from 0.17 to 0.30 standard deviations.

Table 2

Total-Group Data for the Four Pairs of Research Forms To Be Equated

Operational form	Items in each research form	Number of examinees	New form *X*		Reference form *Y*		SMD (*X*–*Y*)
			Mean	SD	Mean	SD	
1	71	13,047	44.82	9.33	47.52	8.90	- 0.30
2	70	14,998	50.60	8.97	47.71	10.15	+ 0.30
3	69	10,251	53.22	6.87	51.81	6.86	+ 0.21
4	63	12,564	45.51	7.45	46.71	6.75	- 0.17

Note. SMD = Standardized mean difference between new form *X* and reference form *Y*.

After constructing a pair of research forms from each of the four operational forms, we had to decide which research form to consider as the new form and which as the reference form. We decided to use the more difficult form as the new form in two cases and as the reference form in the other two cases. Since both examinee populations actually took all the items in both research forms, we also had to decide which population to use as the new form population and which as the reference form population. We decided to create two different data sets for each pair of research forms, using the more able population as the new form population in one data set and as the reference form population in the other data set.

Table 3 shows a statistical comparison of the two research forms and of the two examinee populations in each of the eight data sets used for resampling. The two research forms in Data Set 1A and in Data Set 1B are the same. The two examinee populations in Data Set 1B are exactly reversed from those in Data Set 1A. This pattern is repeated in the other three pairs of resampling data sets.

Table 3

Test Form Difficulty Difference and Examinee Group Ability Difference

Resampling data set (Equating)	New form difficulty (Standardized mean difference in combined population)	New form examinee group ability (Standardized mean difference on full operational form)
1A	Harder (-0.30)	Stronger (+0.21)
1B	Harder (-0.30)	Weaker (-0.21)
2A	Easier (+0.30)	Weaker (-0.14)
2B	Easier (+0.30)	Stronger (+0.14)
3A	Easier (+0.21)	Stronger (+0.22)
3B	Easier (+0.21)	Weaker (-0.22)
4A	Harder (-0.17)	Weaker (-0.30)
4B	Harder (-0.17)	Stronger (+0.30)

Criterion

To evaluate the accuracy of an equating based on samples of examinees, it is necessary to know the population equating that the sample equating is intended to estimate. For each operational form the population for this criterion equating was created by combining the two administration groups (shown in Table 2). Before aggregating the two administration groups, we computed the direct equipercentile equating of the two research forms separately in each group and compared the resulting equating functions. Because the difference between the equated raw scores was smaller than 0.25 across nearly all score points, with all the exceptions below the 1st percentile, we decided that the use of a single criterion equating based on the combined group was justified. Using the scores of all examinees from both administration groups, we performed a direct equipercentile equating of scores on the new form to scores on the reference form. This was the criterion equating for the resampling studies.

Procedure

The resampling studies on each data set included 1,000 replications of the small-sample sampling/equating procedure at each of four specified combinations of sample sizes. The reference-form equating sample in each equating was three times as large as the new-form sample; this is because in operational test equating situations it often is possible to enlarge the reference-form equating sample by combining data from more than one administration of that form. The sample size combinations investigated were (100, 300), (50, 150), (25, 75), and (10, 30). The examinee samples within each replication were selected by simple random sampling without replacement, using SAS PROC SURVEYSELECT. Each replication consisted of the following steps:

1. Select a new-form sample from the new-form population and a reference-form sample from the reference-form population.
2. Equate the new form to the reference form in those samples by each of the small-sample equating methods to be compared.
3. At each new-form raw-score level, for each small-sample equating method, compute the difference between the sample equating and the criterion equating.

After 1,000 replications of the sampling and equating procedure for each equating method, we computed the root-mean-square average of the differences (RMSD) for each small-sample method at each raw-score point. We also computed the RMSD for the identity, which is simply the difference between the identity and the population equating. Computing the RMSD separately at each new-form raw-score value is important when comparing equating methods, because an equating method can be highly accurate in some score regions but highly inaccurate at others.

The RMSD can be decomposed into two orthogonal components: (a) the deviation of the individual small-sample results from their average value, that is, the standard error of equating (SEE); and (b) the deviation of this average small-sample value from the large-group value, that is, the statistical bias of the procedure. The formulas for calculating the bias, SEE, and RMSD at a given score level are as follows.

$$Bias_i = \bar{d}_i = \frac{\sum_{j=1}^{J} d_{ij}}{J} = \frac{\sum_{j=1}^{J}\left[\hat{e}_j(x_i) - e(x_i)\right]}{J}, \tag{1}$$

$$SEE_i = s(d_i) = \sqrt{Var_j\left[d_{ij}\right]} = \sqrt{Var_j\left[\hat{e}_j(x_i) - e(x_i)\right]}, \tag{2}$$

$$RMSD_i = \sqrt{\bar{d}_i^2 + s^2(d_i)}, \tag{3}$$

$$SRMSD_i = \frac{RMSD_i}{SD_y}, \tag{4}$$

where *i* indexes the raw scores in the targeted new form, *j* indexes the replications of the procedure, *J* is the total number of replications (1,000), $\hat{e}_j\left(x_i\right)$ is the equated score for raw score x_i in the *j*th replication of the procedure, $e\left(x_i\right)$ is the equated score for raw score x_i in the criterion equating function, d_i is the difference $\hat{e}_j\left(x_i\right)-e\left(x_i\right)$, and SD_y is the reference-form group standard deviation.

To compare the accuracy of the equating methods in the full examinee population, bias (i.e., systematic error), SEE, and RMSD were each averaged over the new-form raw score distribution, weighting the values at each score level in proportion to the frequency of that raw score in the group of examinees taking the new form. Using f_i to represent the proportional frequency at new-form raw score *i*, the resulting statistics were the weighted root mean squared bias, $\sqrt{\sum_i f_i Bias_i^2}$, the weighted standard error of equating, $\sqrt{\sum_i f_i SEE_i^2}$, and the weighted RMSD, $\sqrt{\sum_i f_i RMSD_i^2}$. We used a root-mean-square averaging procedure to prevent negative bias at one score level from cancelling out positive bias at another. This procedure for computing the average bias also preserves the relationship: bias^2 + SEE^2 = RMSD^2.

Small-Sample Equating Methods

The research reported here compares four methods for estimating an equating transformation from the test scores of samples much smaller than those commonly used in a common-item design. Those four methods are (a) chained equipercentile equating of smoothed distributions, (b) chained linear equating, (c) chained mean equating, and (d) circle-arc estimation method. For comparison, the identity was also used as an estimate of the equating transformation--which would reflect an assumption that the forms were equally difficult at all ability levels.[1]

Chained equipercentile equating of smoothed distributions requires the estimation of as many parameters of the score distributions as are preserved in the smoothing – in these studies, three parameters of each marginal distribution.[2] Chained linear equating requires the estimation of two parameters of each score distribution. Chained mean equating and the circle-arc method both require the estimation of only one parameter of each distribution. Using the identity does not require the estimation of any parameters. Our comparison of these methods is based on the accuracy with which they estimated the direct equipercentile equating in the examinee population formed by combining the two populations from which the equating samples were drawn.

Results

Our investigation included four pairs of test forms to be equated, two assignments of examinee populations to test forms, and four specified sample sizes, for a total of 32 resampling studies. Rather than display the results separately for each of these 32 studies, we have summarized the results in a set of four graphs. There is a single graph for each sample size, showing the results averaged over the eight resampling studies, that is, eight combinations of test-form pair and group assignment. Each graph includes five RMSD curves, one for each small-sample equating method and one for the identity. To create these summary RMSD curves, we averaged the values at selected percentiles of the score distribution, i.e., percentiles of the distribution of scores on the new form in the combined population used to calculate the criterion equating function (these are the groups compared statistically in Table 2). We converted the RMSDs to standard-deviation units of the reference-form examinee group before averaging them across the eight resampling studies. The averaging process was a root-mean-square procedure, squaring the RMSD values for the eight resampling studies, averaging them, and then taking the square root. The vertical scale in the graphs extends from 0.0 to 0.4 SD units; RMSD values of 0.1, 0.2, and 0.3 SD are indicated by dotted horizontal lines. The RMSD curves for each of the 32 separate combinations of test form, equating group, and sample size are shown in the appendix.

Figure 1 shows the RMSD curves for equating in samples of 100 new-form examinees and 300 reference-form examinees, averaged across the four pairs of test forms investigated under the two different group assignments. The most obvious result was that with samples of this size, all equating methods produced much more accurate results than did use of the identity, because the new form and reference form differed in difficulty. At the 50th percentile of the new-form score distribution in the combined population, all the equating methods yielded equally accurate results. At the 10th and 25th percentiles, the two methods that required the estimation of more than one parameter of each score distribution (chained linear and chained equipercentile with 3-moment smoothing) outperformed the two methods based on fewer parameters of each score distribution (chained mean equating and the circle-arc method). Above the 50th percentile, chained mean equating produced much less accurate results than the other three equating methods. The circle-arc method produced the most accurate results at the 95th and 99th percentiles but was less accurate than the other equating methods at the 5th and 10th percentiles.

Figure 2 shows the RMSD curves for equating in samples of 50 new-form examinees and 150 reference-form examinees. The RMSD values were larger than those for the samples of 100 new-form examinees, particularly in the middle of the score distribution, but the comparisons between equating methods were generally similar. Once again the identity produced the least accurate results. From the 10th percentile to the 50th percentile, the differences among the four methods were small. For above-average scores, chained mean equating performed poorly compared to the other methods, but not as poorly as in the larger samples. The circle-arc method produced the most accurate results at the 75th percentile and at all higher percentiles.

Figure 3 shows the RMSD curves for equating in samples of 25 new-form examinees and 75 reference-form examinees. Between the 5th and 90th percentiles, all the equating methods produced more accurate results than the identity. However, above the 90th percentile only the circle-arc method produced more accurate results than the identity. The comparisons among the four methods were similar to those for the larger samples, although the rank ordering of methods at the 95th and 99th percentiles differed. Above the 50th percentile the difference in accuracy between the circle-arc method and the other methods was greater than with the larger samples. Below the 50th percentile chained mean equating produced the most accurate results, followed closely by the circle-arc method.

Figure 4 shows the RMSD curves for equating in samples of 10 new-form examinees and 30 reference-form examinees. With samples of this size, the circle-arc method was the only equating method that produced more accurate results than the identity everywhere from the 1st percentile to the 99th percentile. At the 50th percentile all four equating methods performed similarly. Below the 50th percentile the chained mean and circle-arc methods outperformed the other methods by a wide margin. Above the 50th percentile the circle-arc method was much more accurate than any of the other equating methods. The RMSD values for all the equating methods and for the identity were 0.2 standard deviations or larger throughout the lower half of the score distribution. Below the 25th percentile the RMSDs for the chained linear and chained

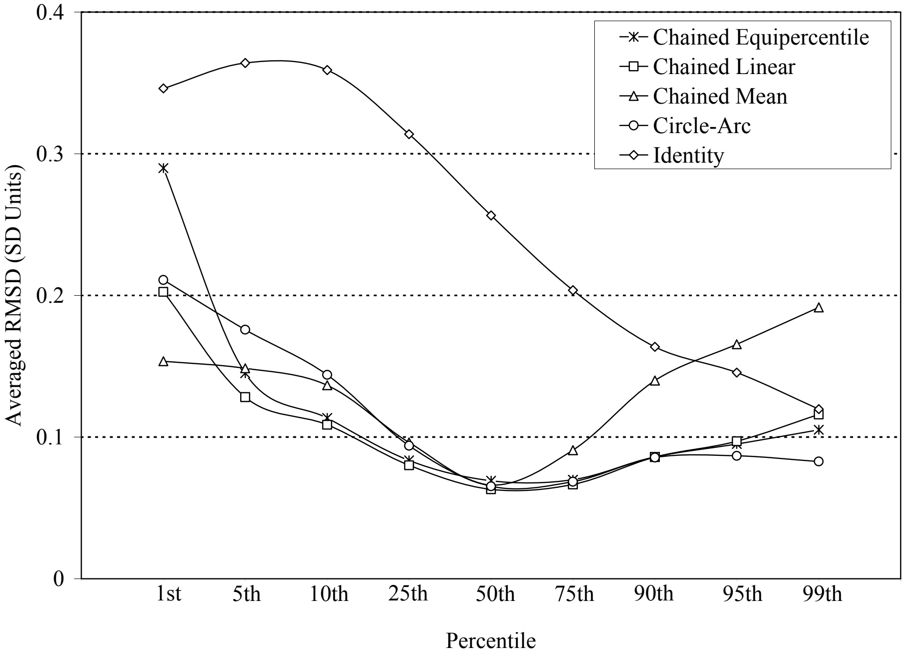


***Figure 1*. Averaged root mean squared difference (RMSD) of equating with samples of 100 new-form examinees and 300 reference-form examinees.**

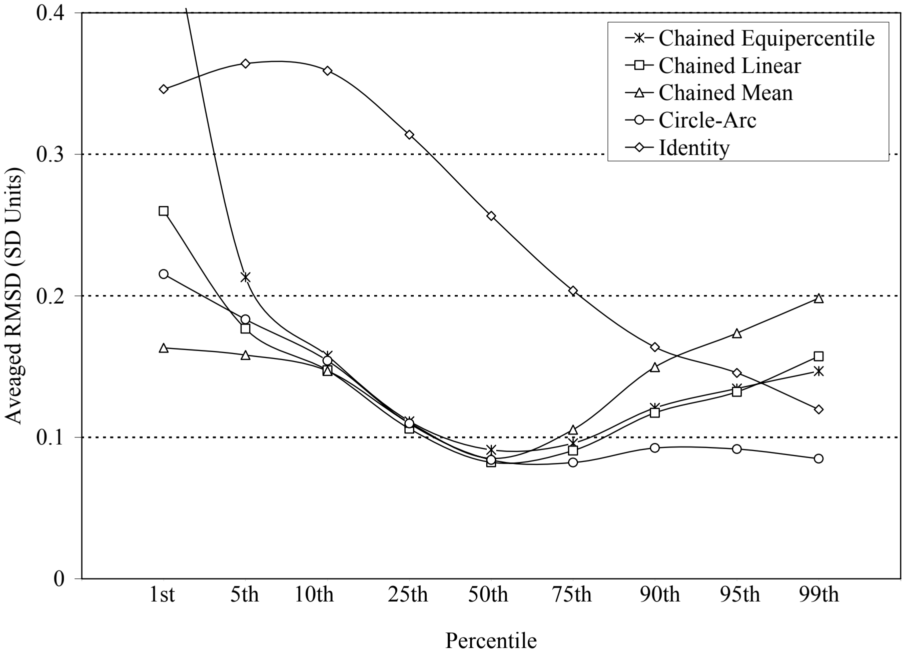


***Figure 2*. Averaged root mean squared difference (RMSD) of equating with samples of 50 new-form examinees and 150 reference-form examinees.**

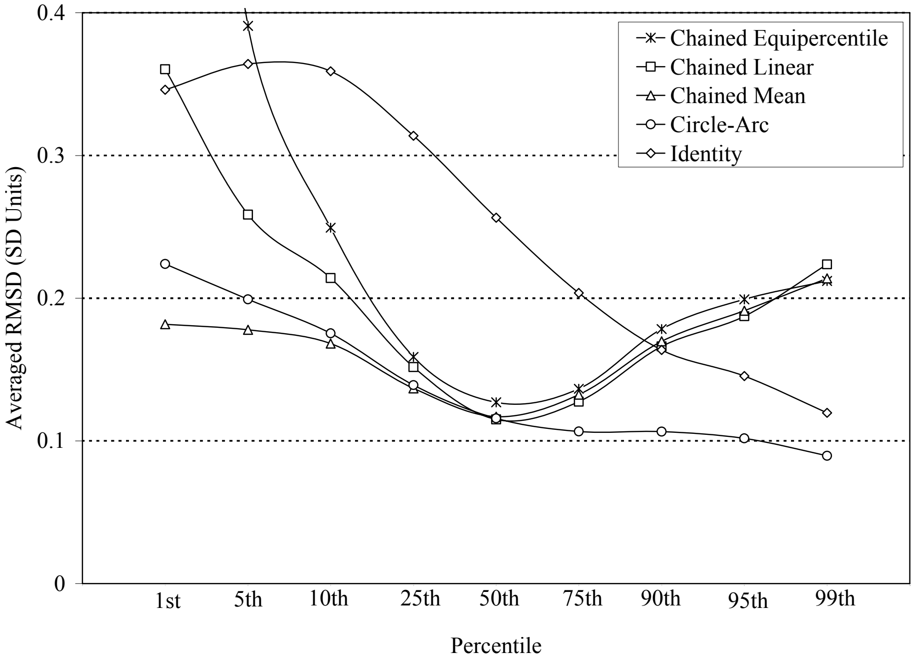


***Figure 3*. Averaged root mean squared difference (RMSD) of equating with samples of 25 new-form examinees and 75 reference-form examinees.**

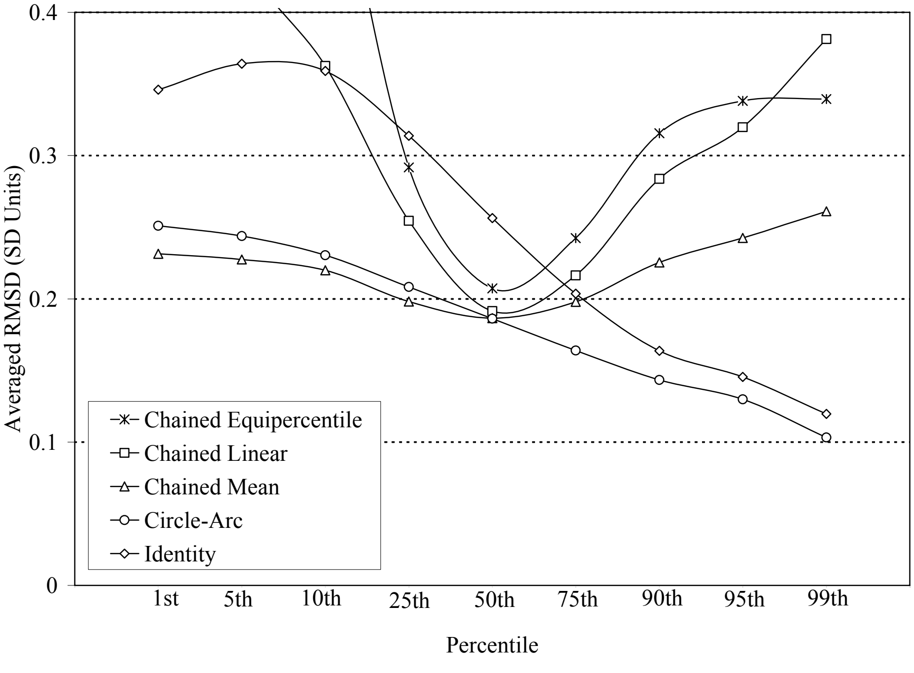


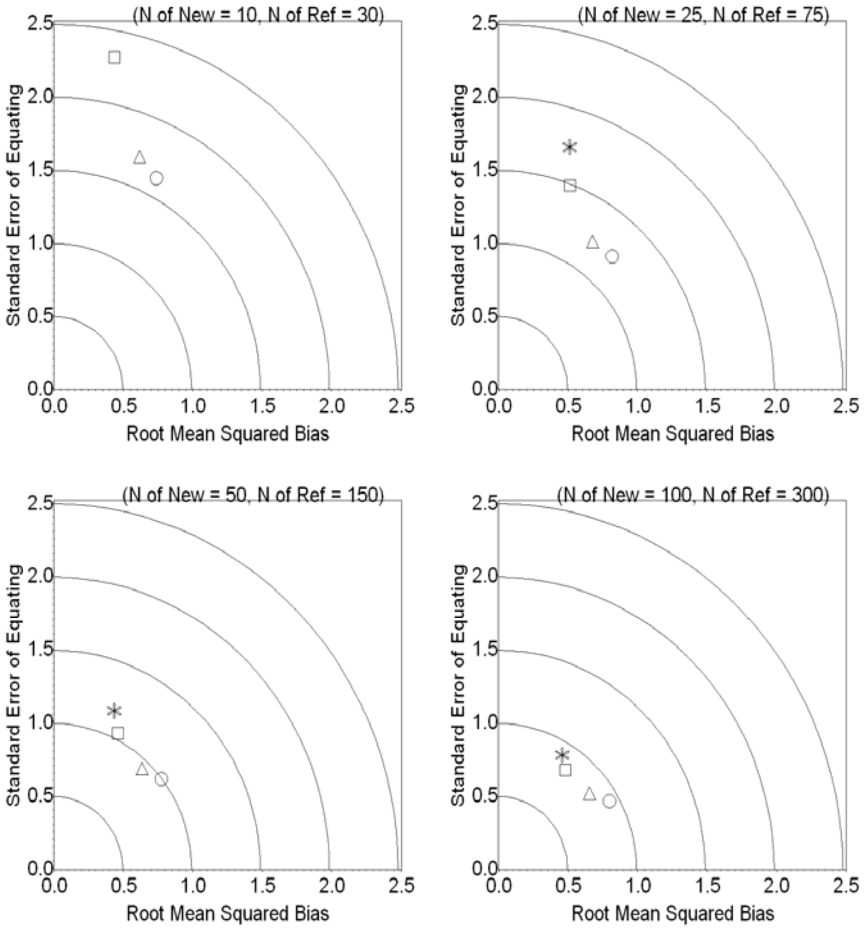
***Figure 4*. Averaged root mean squared difference (RMSD) of equating with samples of 10 new-form examinees and 30 reference-form examinees.**

equipercentile equating methods were greater than 0.4 standard deviations—literally off the chart.

Figures 5A to 8B decompose the overall weighted RMSD into its two orthogonal components: systematic error (bias) and sampling variability. Figures 5A, 6A, 7A, and 8A correspond to the four equatings designated as Equating 1A to Equating 4A in Table 3, respectively. Figures 5B, 6B, 7B, and 8B correspond to the four equatings designated as Equating 1B to Equating 4B in Table 3, respectively. Therefore Figures 5A and 5B refer to the equating of the same two forms, but with the new-form and reference-form populations reversed; likewise for Figures 6A and 6B, and so on. Each figure contains four plots, showing the results for the same pair of test forms equated with different size samples. The horizontal axis in each plot indicates the total (RMS) bias; the vertical axis indicates the standard error in raw-score points. The concentric arcs indicate RMSD values. For example, in the upper left plot of Figure 5A, the total bias of the chained linear method was 0.5 and its SEE was about 2.3. The resulting RMSD of this method ($2.35 = \sqrt{0.5^2 + 2.3^2}$) was located somewhere in the middle between the concentric arc for RMSD=2.0 and the concentric arc for RMSD = 2.5. Also notice that the data point for the chained equipercentile method does not appear in this plot. In some of the figures for the smallest samples, the standard error of equating for a method was so large that the symbol for that method was beyond the range of the vertical scale (0 to 2.5 raw-score points).

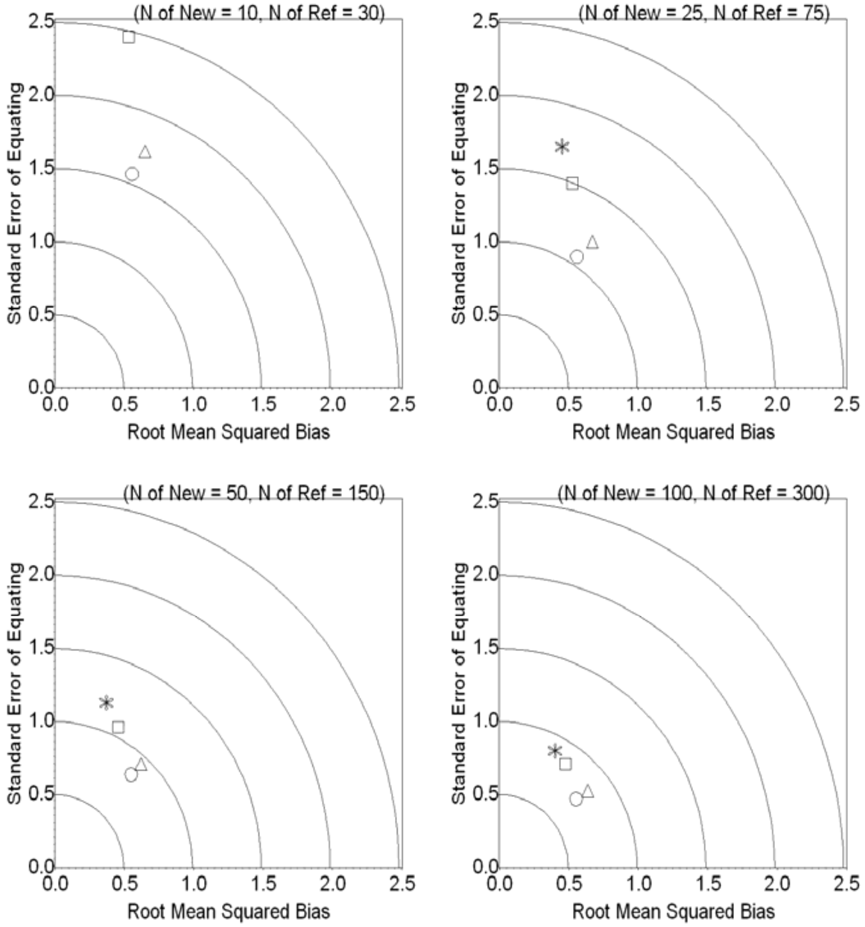
Figure 5A presents the results for the situation in which the new form was harder than the reference form and the new form group was more able than the reference form group. The chained mean and circle-arc methods had slightly more bias but much less sampling error than the chained linear and chained equipercentile methods in the smaller samples, yielding smaller RMSDs. In the largest samples these two effects balanced each other, yielding similar RMSDs for all four equating methods. Figure 5B presents exactly the same information in a situation where equating groups were reversed. The pattern was very similar to that in Figure 5A, except that in Figure 5B the circle-arc method performed better than the other methods across all sample size conditions.

Figure 6A presents the results for the situation in which the new form was easier than the reference form and the new-form group was less able than the reference-form group. The chained mean equating and circle-arc method performed poorly in the larger samples, because of large bias. The circle-arc method had smaller RMSDs than the chained linear and chained equipercentile methods when the equating samples were small. With samples of 50 new-form



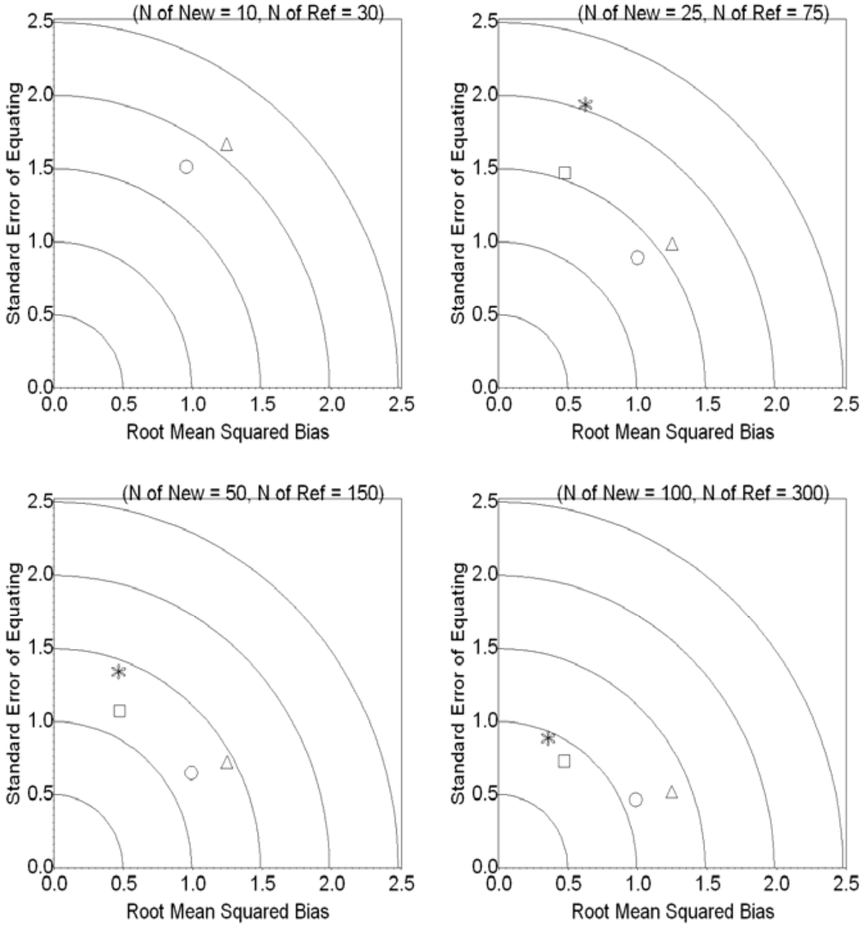
* Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

Figure 5A. **Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating1A (new form more difficult, new-form group more able).**



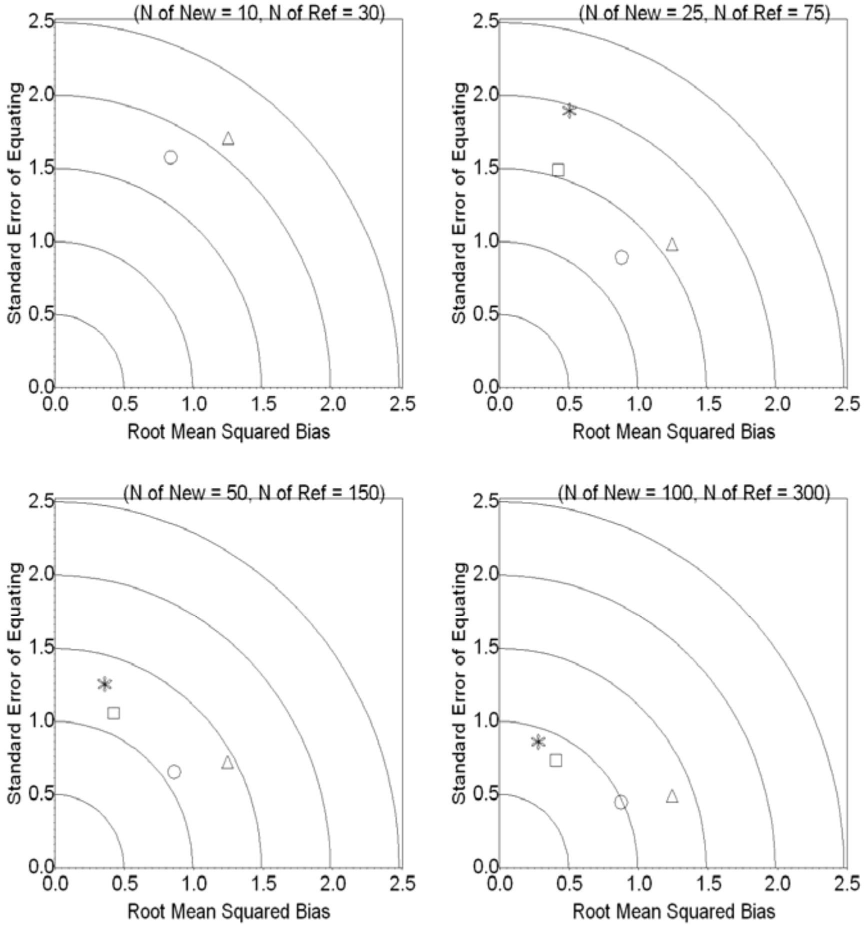
∗ Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

Figure 5B. **Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating1B (new form more difficult, new-form group less able).**



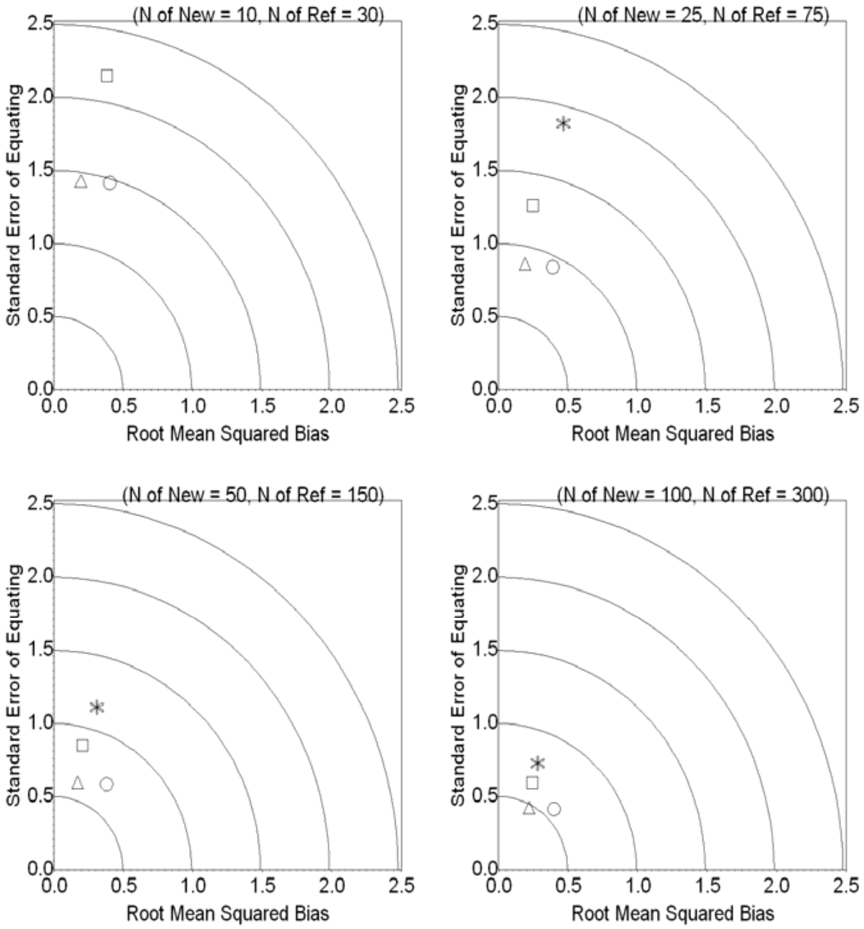
* Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

***Figure 6A*. Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating 2A (new form less difficult, new-form group less able).**



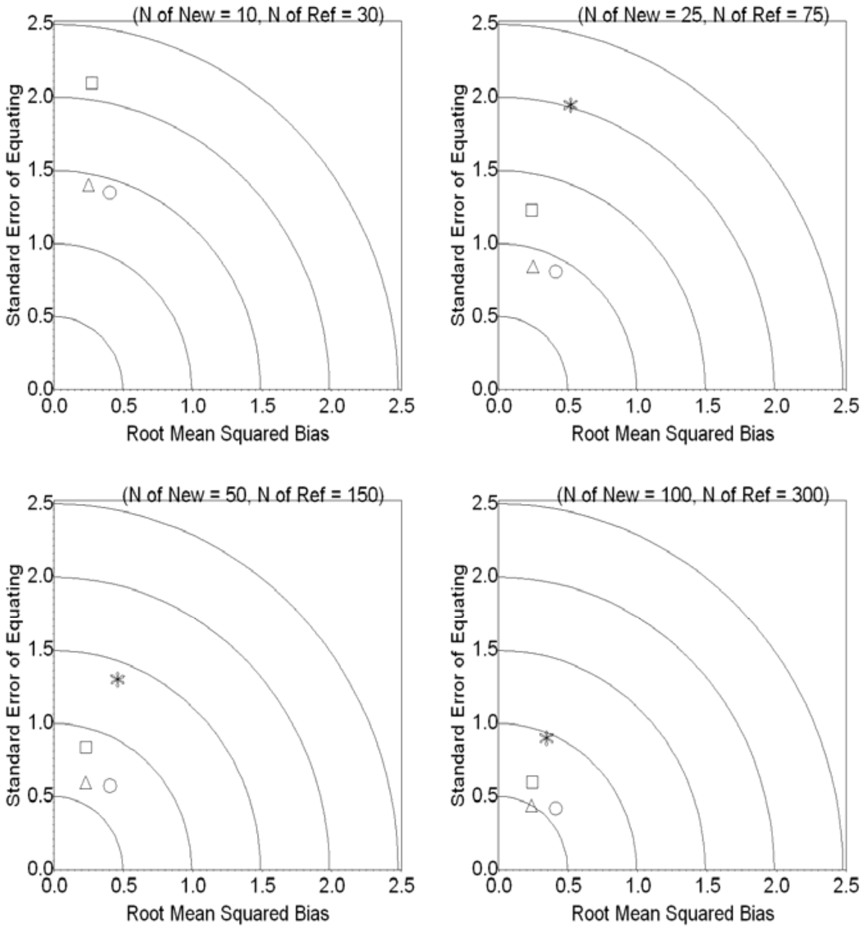
* Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

Figure 6B. **Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating 2B (new form less difficult, new-form group more able).**



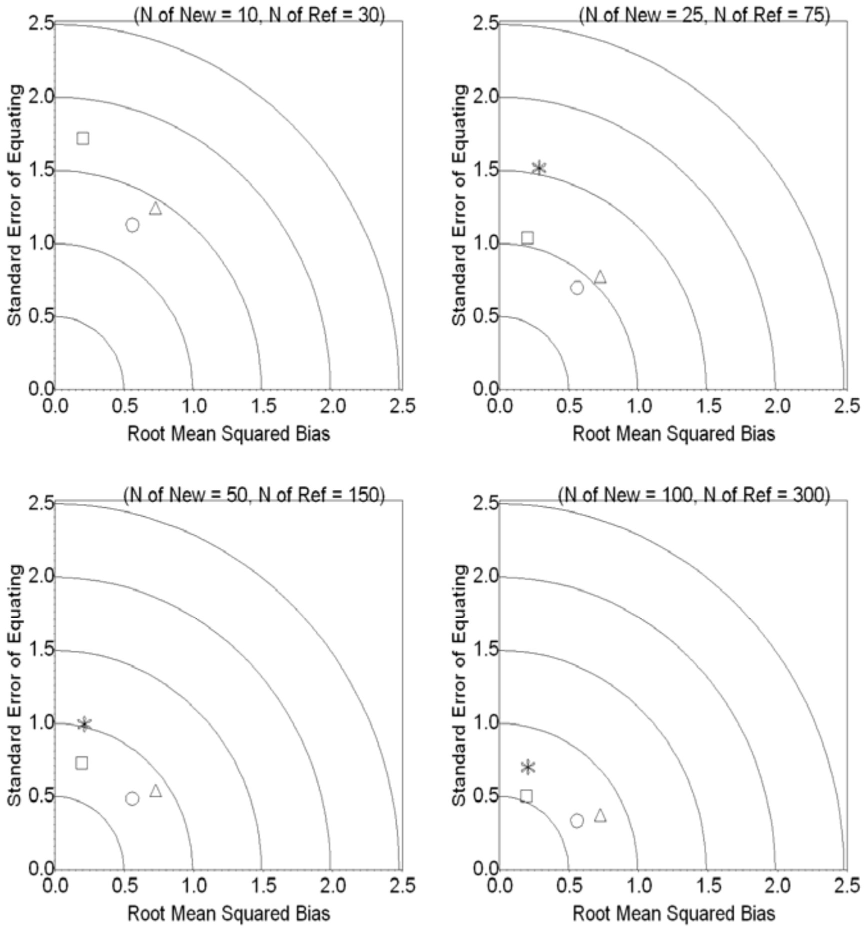
* Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

Figure 7A. **Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating 3A (new form less difficult, new-form group more able).**



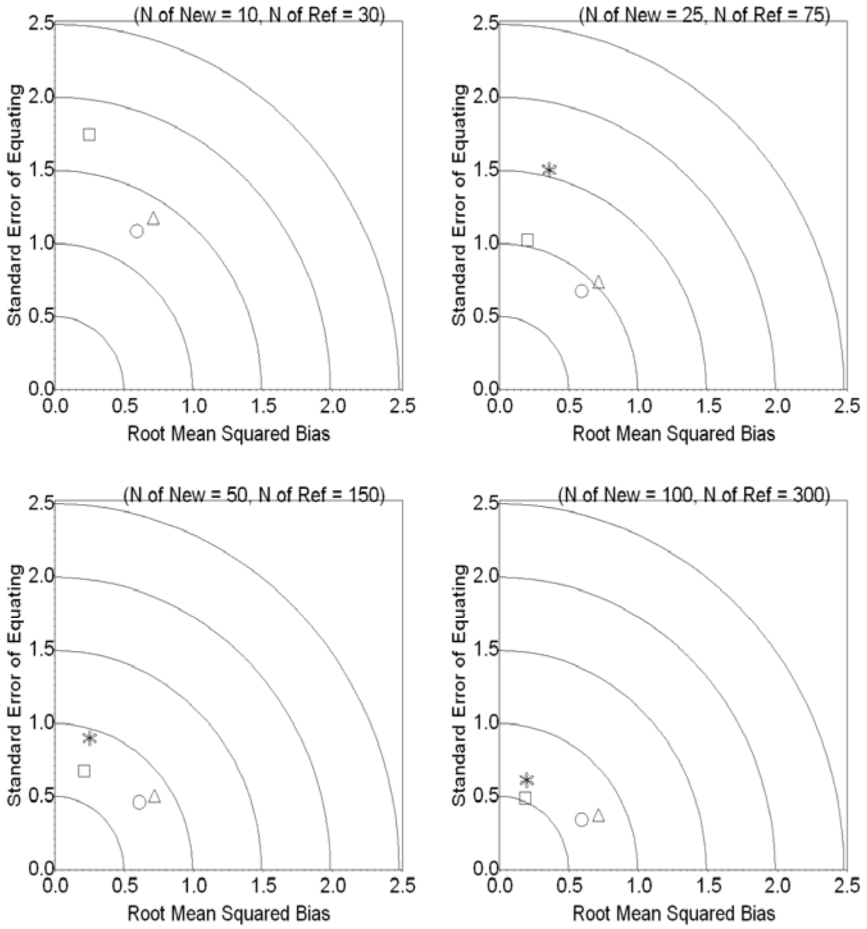
* Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

Figure 7B. **Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating 3B (new form less difficult, new-form group less able).**



* Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

Figure 8A. **Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating 4A (new form more difficult, new-form group less able).**



* Chained Equipercentile □ Chained Linear Δ Chained Mean ○ Circle Arc ___RMSD

Figure 8B. **Plots of the weighted averaged root mean squared difference (RMSD) as a function of the weighted average root mean squared bias and error: Equating 4B (new form more difficult, new-form group more able).**

and 150 reference-form examinees, the larger sampling error of the chained linear and chained equipercentile methods approximately balanced the larger systematic error of the chained mean and circle-arc methods. With the largest samples, the smaller sampling error of the chained linear and chained equipercentile methods enabled them to outperform the circle-arc and chained linear methods. As shown in Figure 6B, even though equating groups were reversed, a very similar pattern emerged, leading to almost identical conclusions regarding the four equating methods with most sample sizes.

Figure 7A presents the results for the situation in which the new form was easier than the reference form and the new-form group was more able than the reference-form group. In this case there was very little systematic error in the chained mean and circle-arc methods, enabling them to outperform the other methods for all sample sizes investigated. Chained mean equating showed slightly less systematic error than the circle-arc method for this pair of forms; the amount of sampling error in those two methods was equal. Figure 7B shows the same pattern for this pair of forms with the two populations reversed.

Figure 8A presents the results for the situation in which the new form was harder than the reference form and the new-form group was less able than the reference-form group. The pattern was similar to that of Figure 6A, but with smaller standard errors for the chained linear and chained equipercentile methods and smaller bias for all four methods. Figure 8B shows that these results did not change when the two populations were interchanged.

Discussion

Accurate equating requires representative samples of examinees. The possibility of obtaining representative samples from a population is greatly diminished when only small samples are available. A lack of sufficient data is likely to affect both test assembly (because of, for example, no pretest item statistics and unstable item statistics derived from the small number of examinees) and equating processes. This means that forms for low-volume tests are more likely to be non-parallel than are forms developed from tests for which ample data are usually available for test assembly. When test forms vary in difficulty, their distributions in a population tend to be unequally skewed. The difference in the shape of the distribution leads to a curvilinear equating transformation. The circle-arc method (Livingston & Kim, 2008, 2009) provides a way to use small-sample data to estimate a curvilinear equating relationship.

We conducted the present study to investigate the effectiveness of the circle-arc method under the common-item design. In this investigation of four methods for equating in the common-item design, the circle-arc method tended to produce more accurate results than the other methods with small samples – 25 or fewer examinees for the new form and 75 or fewer for the reference form. The smaller the samples, the greater the advantage in accuracy of the circle-arc method over the other methods. With samples four times as large – 100 for the new form and 300 for the reference form – both chained linear and chained equipercentile equating outperformed the circle-arc method, particularly at the 25th percentile and below. The circle-arc method was particularly accurate for equating at high score levels. All four methods were about equally accurate at the 50th percentile. As scores increased beyond the 75th percentile, the three other equating methods all produced less accurate results, whereas the circle-arc method became more accurate. For scores at or below the 25th percentile, both the circle-arc method and chained mean equating produced more accurate results than the other methods for samples of 25 or fewer examinees taking the new form and 75 or fewer taking the reference form.

The circle-arc method not only outperformed the other methods with very small samples; at the highest score levels it also produced the most accurate equating results even with as many as 100 examinees taking the new form and 300 taking the reference form. The circle-arc method equates the maximum possible scores on the new and reference forms. This feature yields an estimated equating curve that resembles those typically produced by equipercentile equating. Like mean equating, this method requires the estimation of only a single point on the equating curve, so this feature enables the method to perform well with small amounts of data. But unlike mean equating, it forces the estimated equating function to curve in the same way that equipercentile equating functions usually curve. Because of those features, the overall performance of the circle-arc method was superior to chained mean equating even with samples of 25 new-form and 75 reference-form examinees.

Chained mean equating performed well for average and below-average scores, but not for high scores. Its main limitation appears to be an inability to model a curvilinear equating relationship. Chained linear and chained equipercentile equating both performed poorly for scores below the 25th percentile or above the 75th percentile, especially with samples of 25 new-form and 75 reference-form examinees (or fewer). The problem with chained equipercentile equating is that the percentile ranks of scores in those regions are not accurately estimated in small samples. Chained linear equating cannot estimate a curvilinear relationship, and the slope of the conversion often is not estimated accurately with small samples. The use of the identity was not a good substitute for equating with the four pair of reassembled forms, because those forms were constructed specifically to differ in difficulty in order to produce a situation in which equating is necessary. With small samples of 10 new-form and 30 reference-form examinees, chained linear equating and chained equipercentile equating were so inaccurate that they were no better than assuming equal difficulty for test forms that clearly were not equally difficult. Even with samples of that size, however, chained mean equating was more accurate than the identity below the 75th percentile, and the circle-arc estimate was more accurate than the identity throughout the entire score range.

An interesting question arises from the current findings: Why did the circle-arc method work better than the other methods at the high end of the score distribution but not at the low end? One possible explanation is that even when forms differ in difficulty, the highest score with a nonzero frequency in the population tends to be the same (or nearly the same) on both forms. Therefore the equipercentile equating relationship in the population nearly always comes very close to the intersection of the maximum possible scores, which is the upper end-point in the circle-arc method. But the lowest score with a nonzero frequency in the population may be quite a bit higher on an easy form than on a hard one. Therefore the equipercentile equating relationship in the population may not come close to the intersection of chance scores (or any other prespecified minimum scores) on the two forms. For three of the four pairs of forms in this study, the equipercentile equating curve in the full, combined population did not come close to the intersection of chance scores. For two of the four pairs of forms, it was not close to any point on the identity line in the lower portion of the score range.

As with any other approach to small-sample equating, however, the circle-arc method has limitations. Mathematically, it does not meet the requirement that an equating method must be symmetric, although there is a symmetric version available (Livingston & Kim, 2008). Therefore, applied to population data, it is not truly an equating method. However, its purpose is to estimate the equipercentile equating in the population (which is symmetric) from small-sample data, and the estimates it produced in these studies were more accurate than those of the other methods investigated.

A more practical limitation of the circle-arc method, revealed in these studies, is that it was not nearly as accurate at low score levels as it was at high score levels. The method requires the user to specify the lowest meaningful score on each form, and that decision is somewhat arbitrary. Livingston and Kim (2008, 2009) suggested the use of the chance score on a multiple-choice test as the lowest meaningful score. However, for some multiple-choice tests, a lower score – possibly even zero – might be a better choice. As the current findings indicate, this limitation can be mitigated by using the circle-arc method to estimate the conversion above the new-form mean and using mean equating to estimate the conversion below the new-form mean.

In the present study all the equating situations used data from populations where the distributions of scores were negatively skewed. Would the relative accuracy of the different equating methods be the same in a population where the scores were positively skewed? To answer this question we repeated the resampling studies with populations consisting entirely of "repeaters" – examinees who tend to be of lower ability than first-time examinees for those particular tests, using only two of the four pairs of forms. Repeaters tend to be of lower ability than first-time examinees, because most examinees whose scores are sufficient to earn them a license to teach do not take the test again. The mean differences between these two distinct subgroups, repeaters and first-timers, were larger than one standard deviation. The score distributions based solely on repeaters were positively skewed. The results of these studies were similar to those of the studies using the full examinee populations. One noticeable difference was that the chained mean method performed as well as the circle-arc method for above-average scores ranging from the 75th percentile to the 90th percentile. We then repeated the resampling studies once again, using the repeater populations and setting the lower end-point of the circle-arc method curve at zero, rather than at the chance score. With this change, the circle-arc method performed as well as the chained mean method for below-average scores ranging from the 5th percentile to the 25th percentile, and performed better than the other methods at the high score region as well. A more sophisticated study would be required to generalize this result to other situations, for example, when the distributions are positively skewed.

The purpose of the circle-arc method is to estimate the equipercentile equating in the population (which is symmetric) from small-sample data. The results of these studies, obtained with a common-item equating design, were similar to those obtained previously from similar studies using an equivalent-groups equating design (Livingston & Kim, in press). Both resampling studies support the use of the circle-arc method in practice as an alternative approach for small-sample nonlinear equating. The circle-arc method may be preferable to linear equating (e.g., chained mean or chained linear) when test forms vary in difficulty and samples are too small for equipercentile equating. However, we do not propose the circle-arc method as a solution or methodological fix for problems caused by poor data collection practices. When small samples are unavoidable, this method could be utilized along with other strategies, such as inclusion of a reasonably large number of common items in a common-item design. This would be a good strategy because multiple-choice item tests with a high percentage of overlap and a high correlation between anchor and total test scores reduce equating error and bias.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: ETS.

Kim, S., Livingston, S., & Lewis, C. (2008). *Investigating the effectiveness of collateral information on small sample equating* (ETS Research Rep. No. RR-08-52). Princeton, NJ: ETS.

Kim, S., Livingston, S., & Lewis, C. (in press). Collateral information for equating in small samples: A preliminary investigation. *Applied Measurement in Education.*

Kim, S., von Davier, A. A., & Haberman, S. (2007, April). *Investigating the effectiveness of a synthetic linking function on small sample equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small sample equating using a synthetic linking function. *Journal of Educational Measurement*, *45,* 325-342.

Kim, S., von Davier, A. A., & Haberman, S. (in press). Practical application of a synthetic linking function on small-sample equating. *Applied Measurement in Education.*

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

Livingston, S. A. (1993). Small sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*, 23-39.

Livingston, S. A., & Kim, S. (2008). *Small-sample equating by the circle-arc method* (ETS Research Rep. No. RR-08-39). Princeton, NJ: ETS.

Livingston, S. A., & Kim, S. (in press). Random-group equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46,* 330-343.

Livingston, S. A., & Lewis, C. (2009). *Small sample equating with prior information* (ETS Research Rep. No. RR-09-25). Princeton, NJ: ETS.

Parshall, C. G., Houghton, D. B. P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*, 37-54.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42,* 309–330.

Notes

[1] Some authors have recommended using the identity as the equating transformation whenever the available samples of test takers are below a prespecified size (Kolen & Brennan, 2004, pp. 289–290; Skaggs, 2005, p. 309).

[2] The information in the bivariate moments is not used in chained equating, which operates only on the marginal distributions.

Appendix

Table A1

Statistical Comparisons for the Large-Group Equatings: Case A

	Equating 1		Equating 2		Equating 3		Equating 4	
Form	New	Ref	New	Ref	New	Ref	New	Ref
Number of items	71	71	70	70	69	69	63	63
Number of anchor items	25	25	23	23	28	28	29	29
% of anchor items	35%	35%	33%	33%	41%	41%	46%	46%
Number of examinees	6,580	6,467	6,180	8,818	5,475	4,776	7,130	5,434
Total score mean	45.79	46.66	49.87	48.27	53.88	51.00	44.60	47.77
Total score SD	(9.00)	(9.16)	(9.29)	(9.87)	(6.67)	(7.04)	(7.60)	(6.40)
Anchor score mean	16.47	15.90	15.26	15.74	21.79	21.23	20.77	21.74
Anchor score SD	(3.47)	(3.70)	(3.84)	(3.65)	(2.99)	(3.12)	(3.79)	(3.53)
Correlation of total & anchor	.88	.90	.91	.90	.86	.85	.90	.90

Table A2

Statistical Comparisons for the Large-Group Equatings: Case B

	Equating 1		Equating 2		Equating 3		Equating 4	
Form	New	Ref	New	Ref	New	Ref	New	Ref
Number of items	71	71	70	70	69	69	63	63
Number of anchor items	25	25	23	23	28	28	29	29
% of anchor items	35%	35%	33%	33%	41%	41%	46%	46%
Number of examinees	6,467	6,580	8,818	6,180	4,776	5,475	5,434	7,130
Total score mean	43.82	48.37	51.11	46.91	52.47	52.51	46.71	45.90
Total score SD	(9.55)	(8.56)	(8.70)	(10.49)	(7.02)	(6.62)	(7.07)	(6.90)
Anchor score mean	15.90	16.47	15.74	15.26	21.23	21.79	21.74	20.77
Anchor score SD	(3.70)	(3.47)	(3.65)	(3.84)	(3.12)	(2.99)	(3.53)	(3.79)
Correlation of total & anchor	.89	.87	.90	.91	.86	.85	.90	.91

The following graphs are organized into four sets of nine graphs each. In each set, the first graph shows the criterion equating compared with the identity. The other eight graphs in a set show the results of the resampling study at each of four specified sample sizes of examinees taking each form under the two assignments of equating groups as shown in Tables A1 and A2. The four specified sample sizes are: (100, 300), (50, 150), (25, 75) or (10, 30).

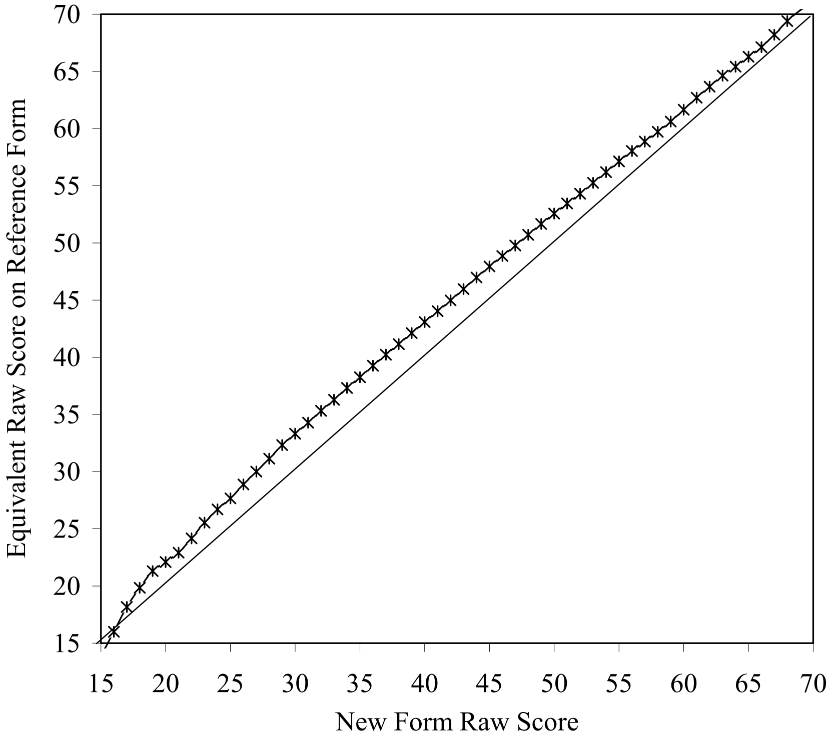


Figure A1. **Criterion equating for Research Form 1.**

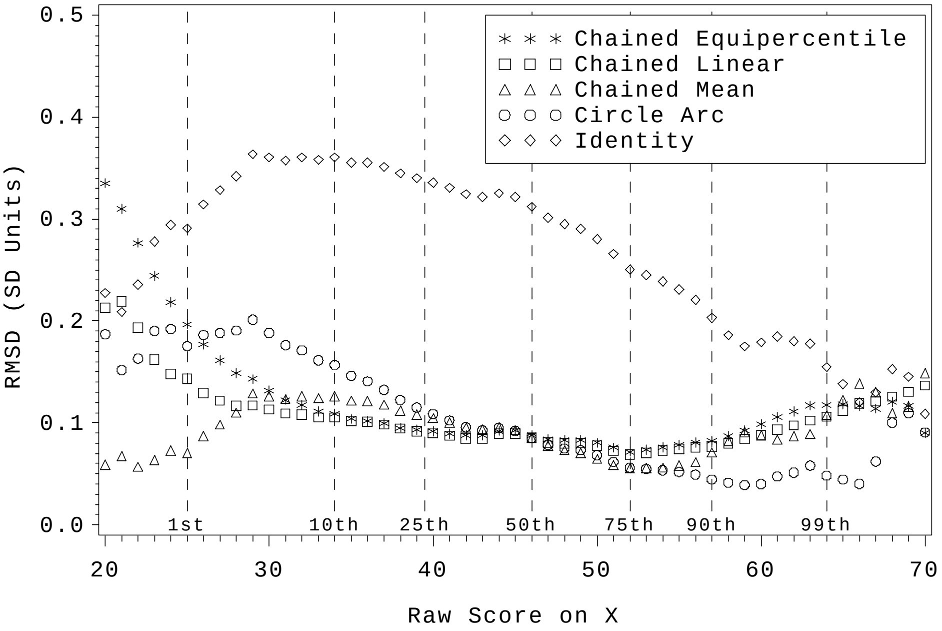


Figure A2. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 1A.**

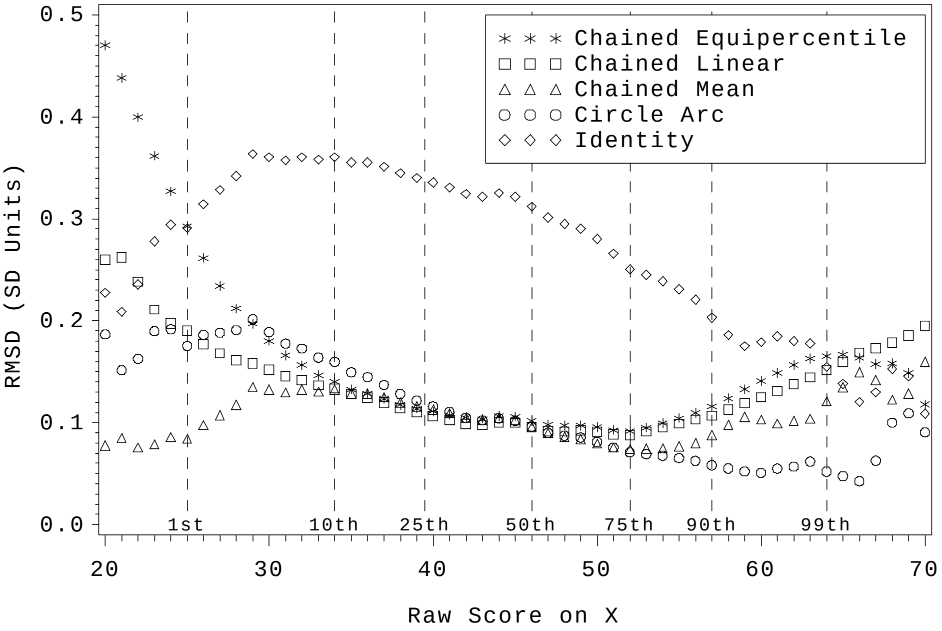


Figure A3. **Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 1A.**

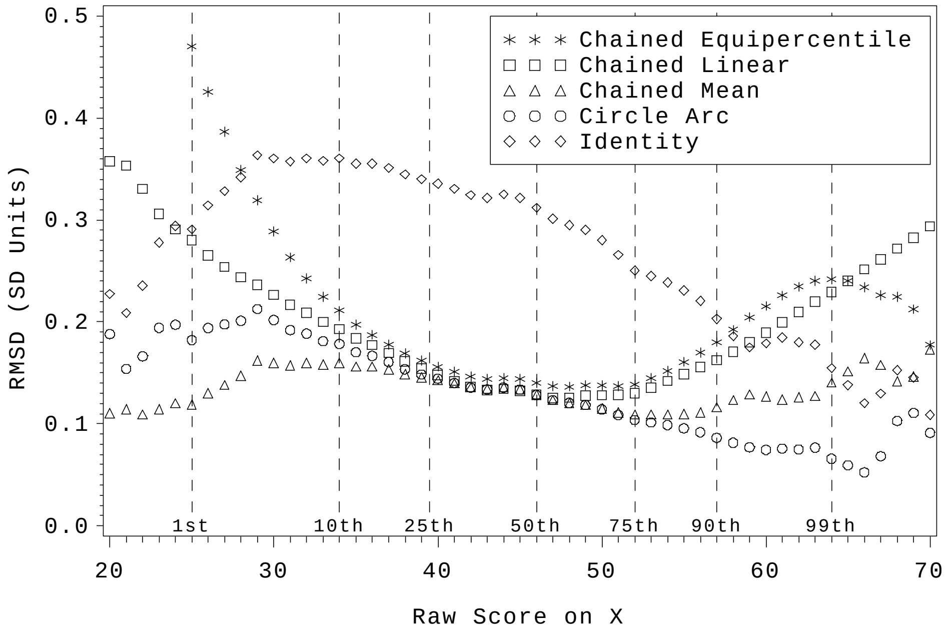


Figure A4. **Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 1A.**

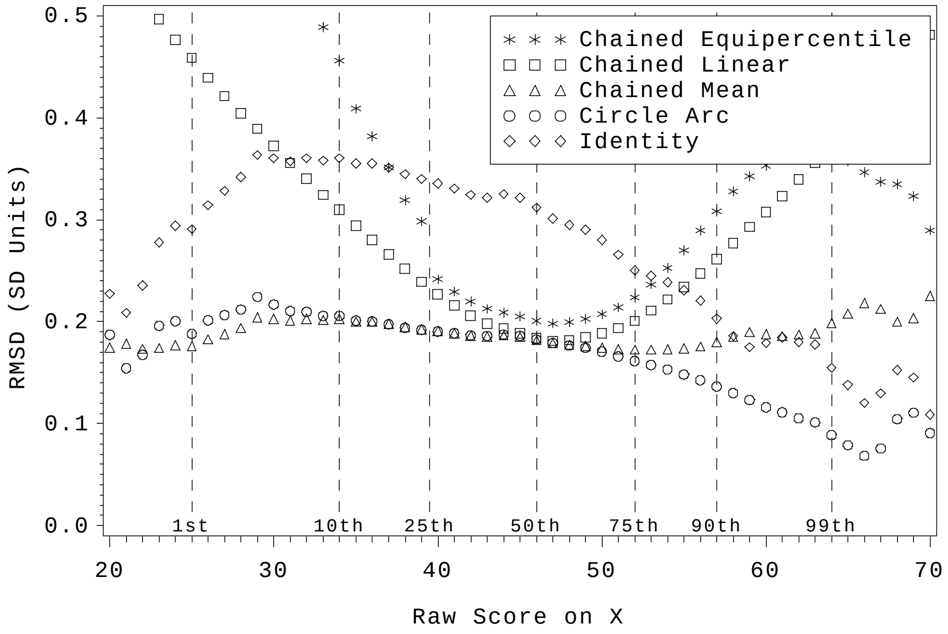


Figure A5. **Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 1A.**

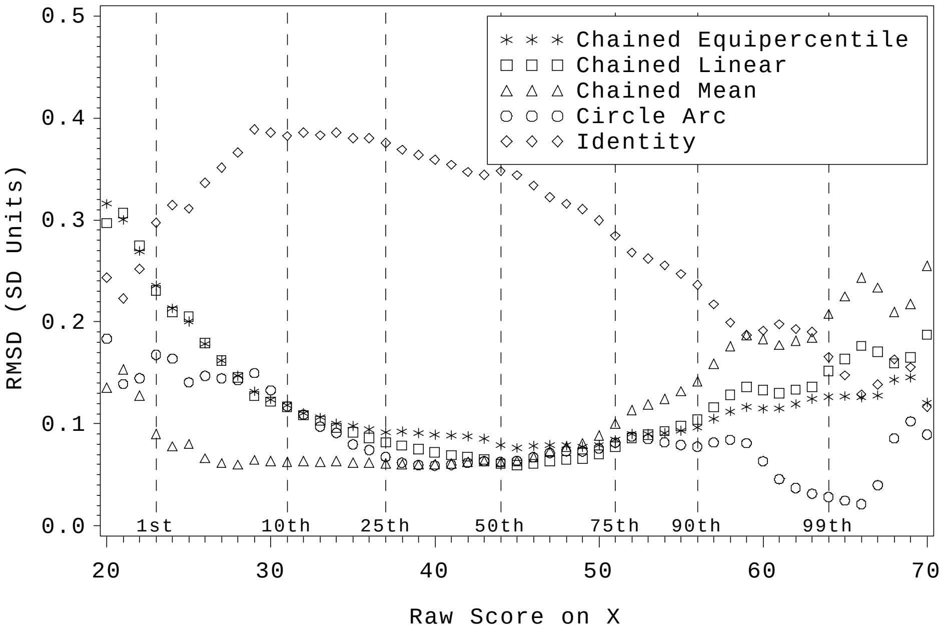


Figure A6. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 1B.**

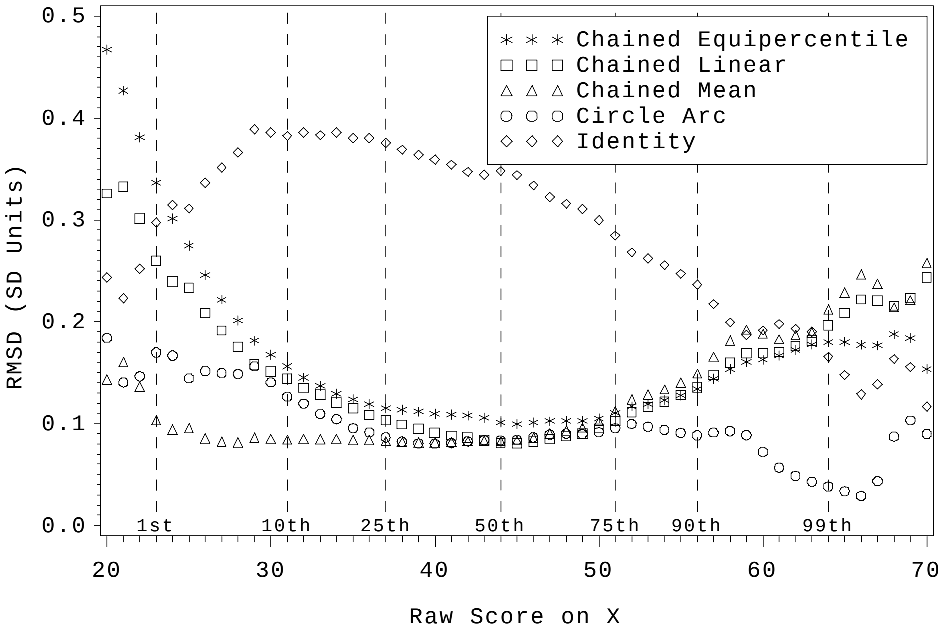


Figure A7. **Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 1B.**

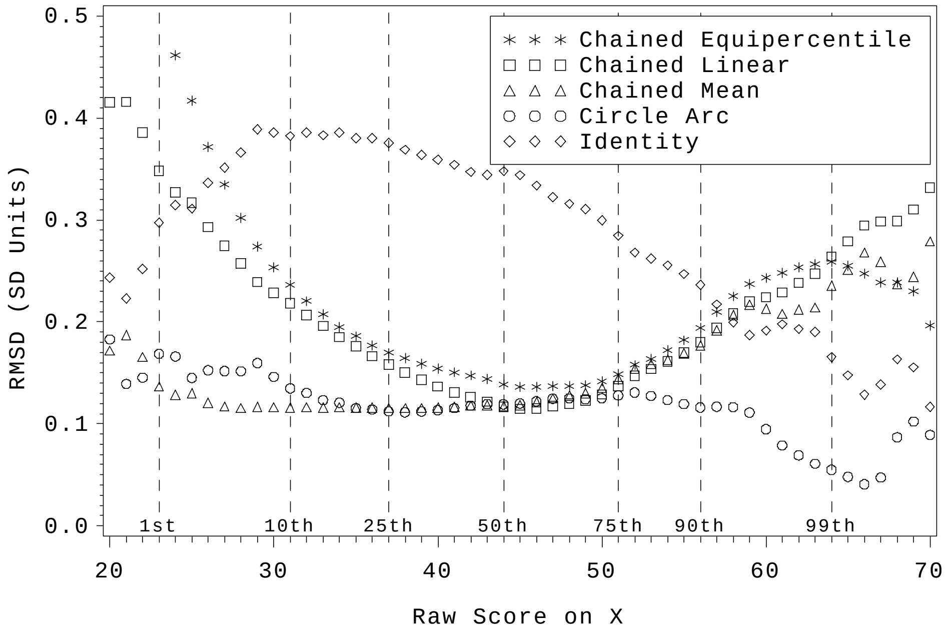


Figure A8. **Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 1B.**

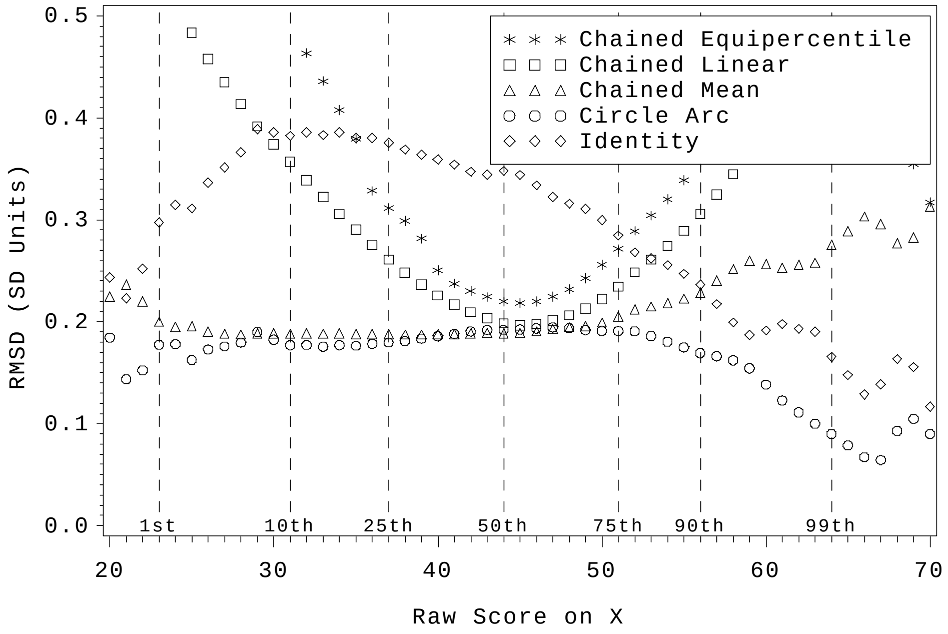


Figure A9. **Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 1B.**

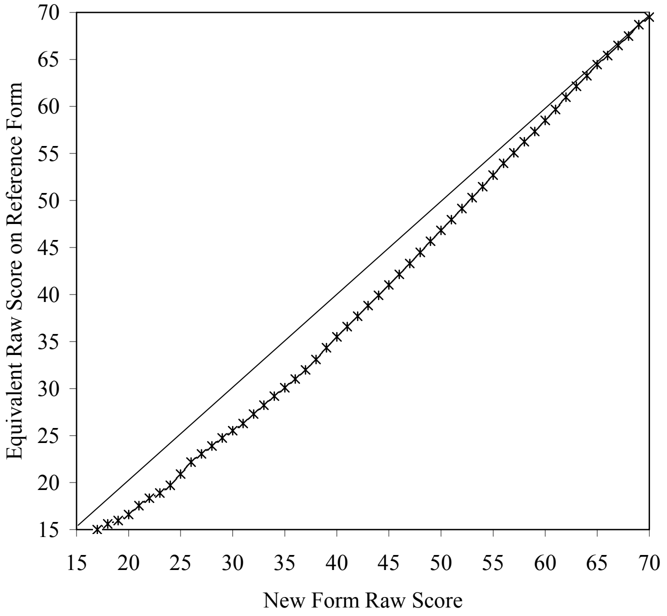


Figure A10. **Criterion equating for Research Form 2.**

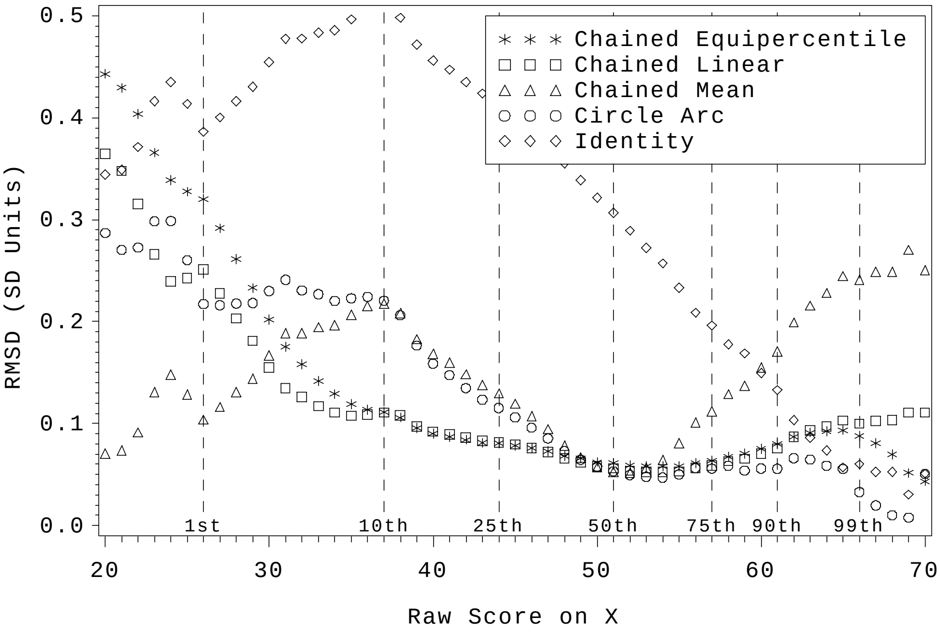


Figure A11. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 2A.**

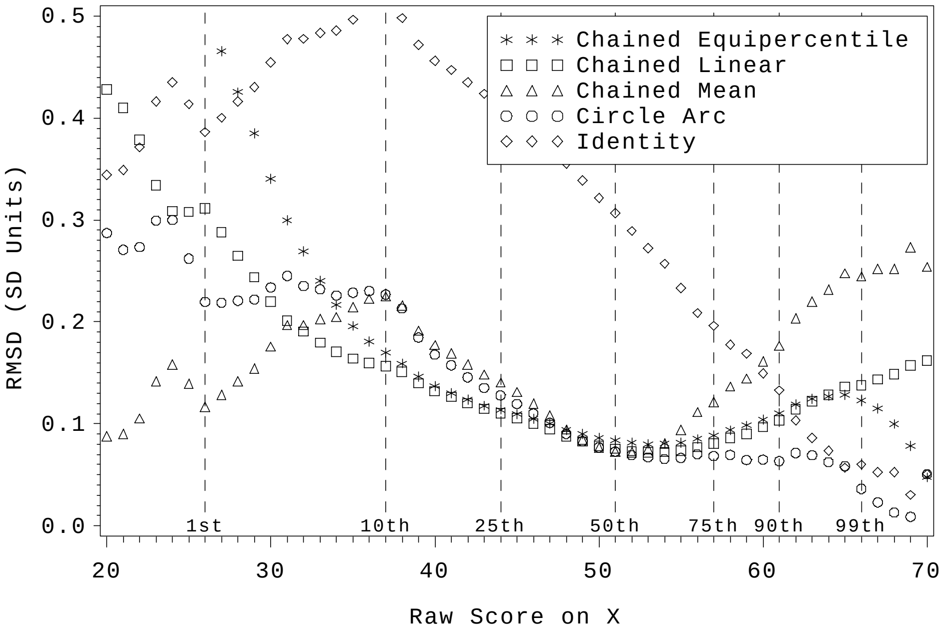


Figure A12. **Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 2A.**

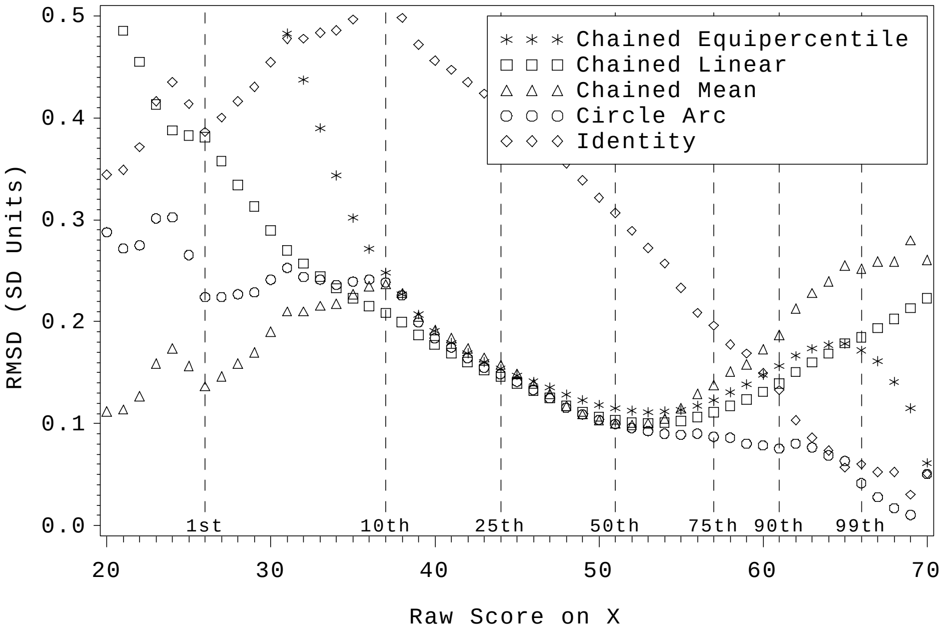


Figure A13. **Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 2A.**

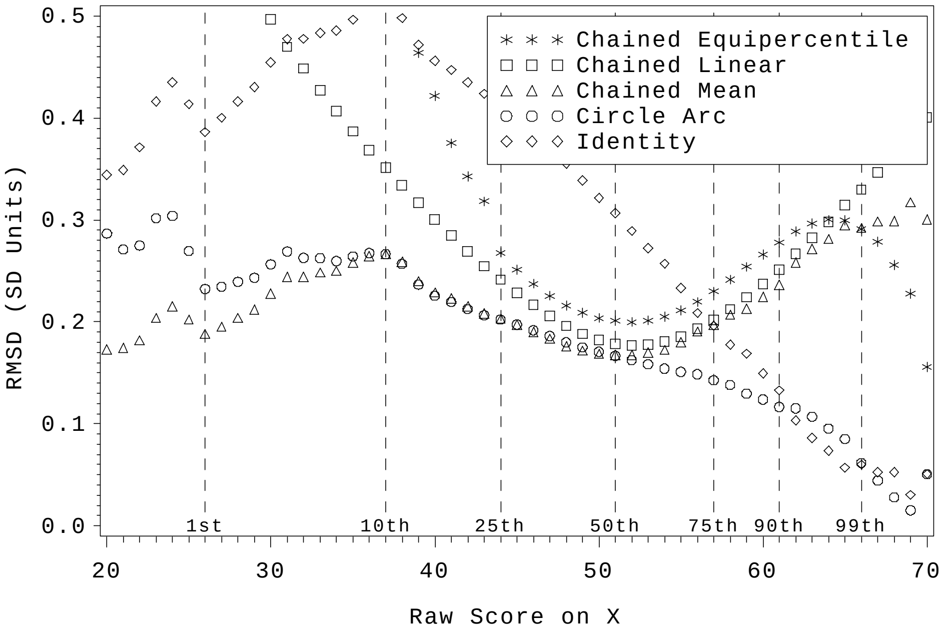


Figure A14. **Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 2A.**

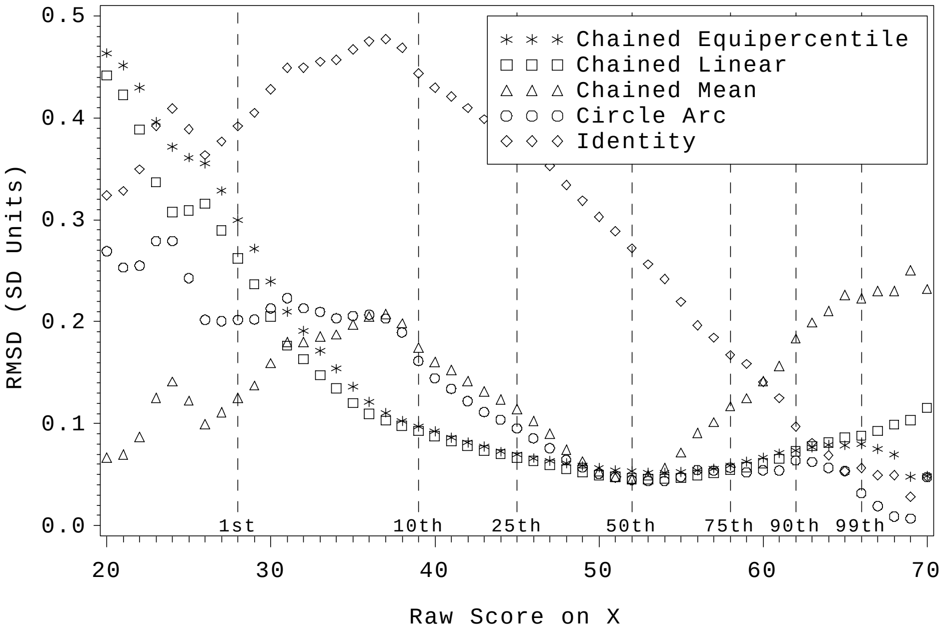


Figure A15. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 2B.**

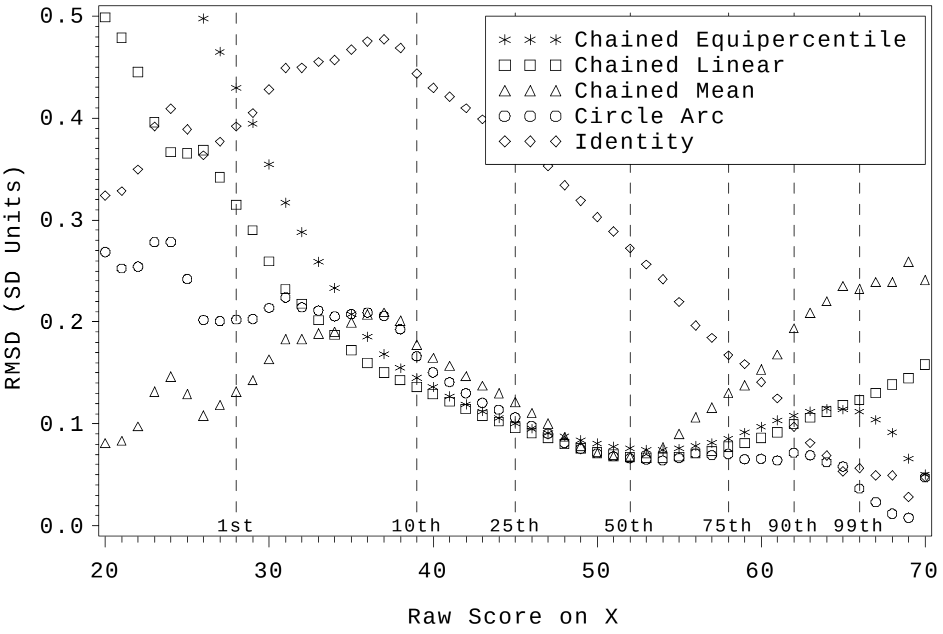


***Figure A16*. Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 2B.**

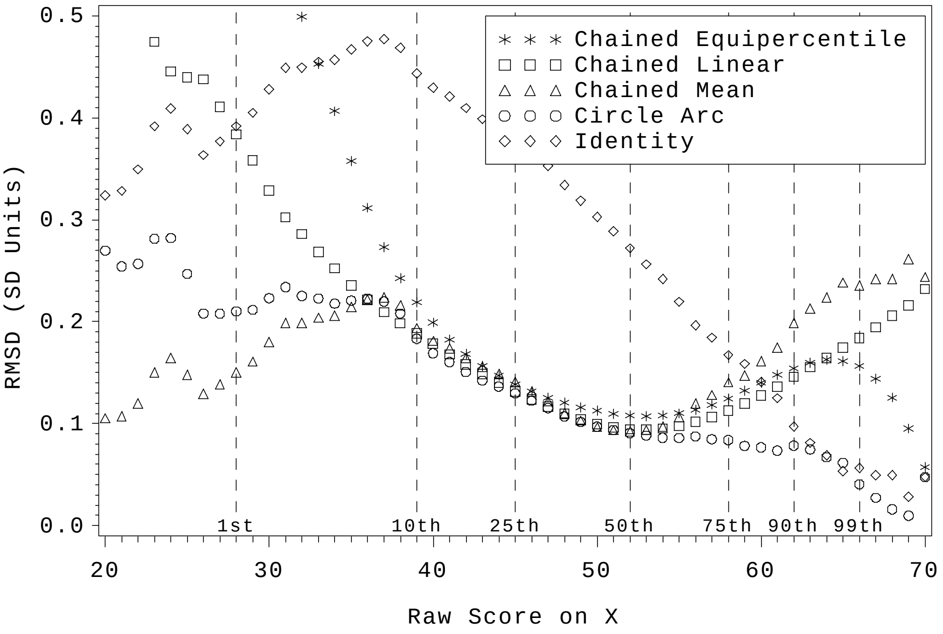


***Figure A17*. Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 2B.**

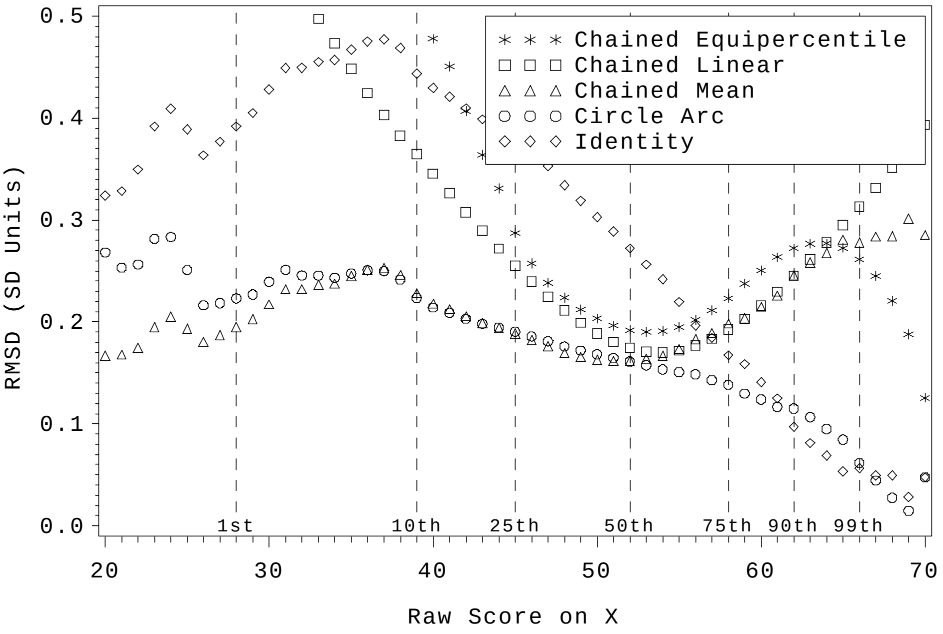


Figure A18. **Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 2B.**

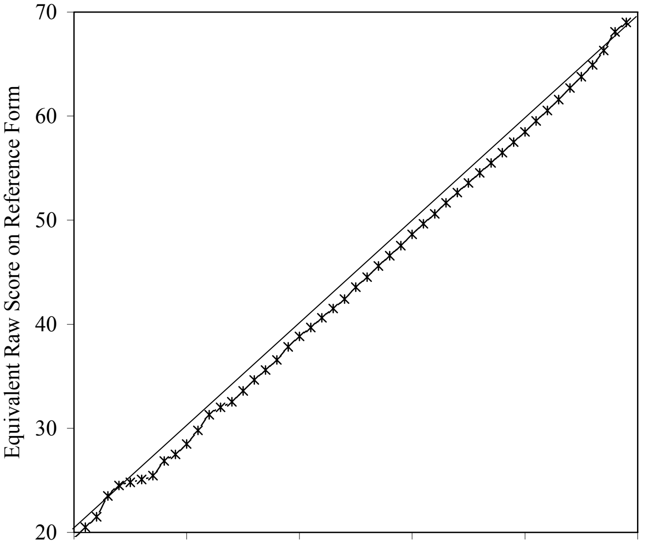


Figure A19. **Criterion equating for Research Form 3.**

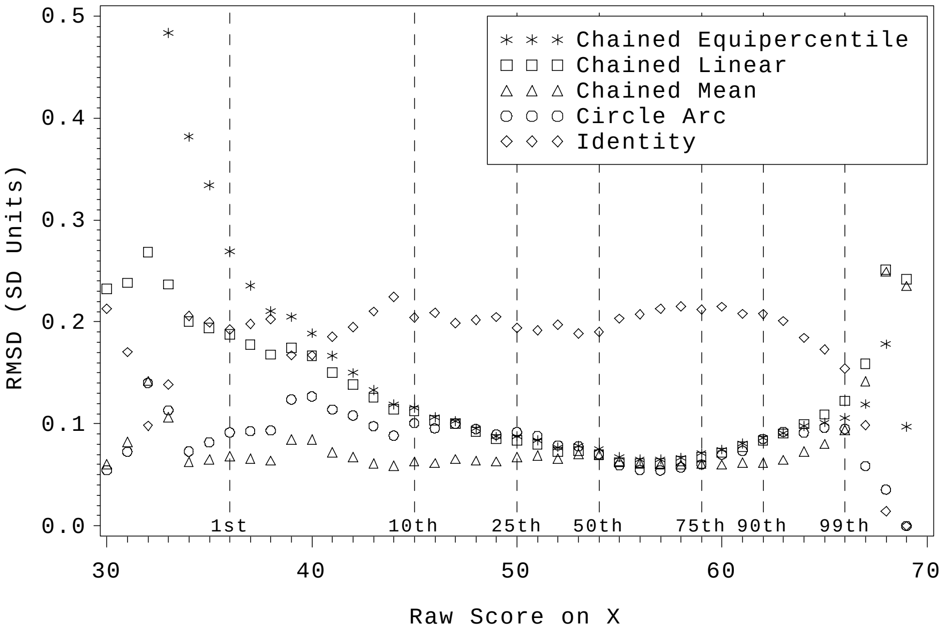


Figure A20. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 3A.**

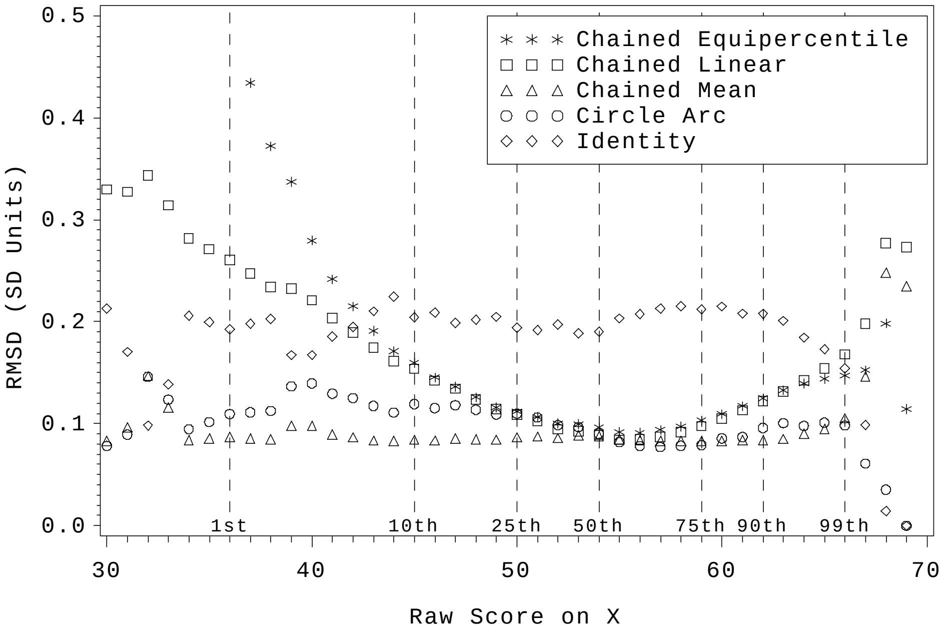


Figure A21. **Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 3A.**

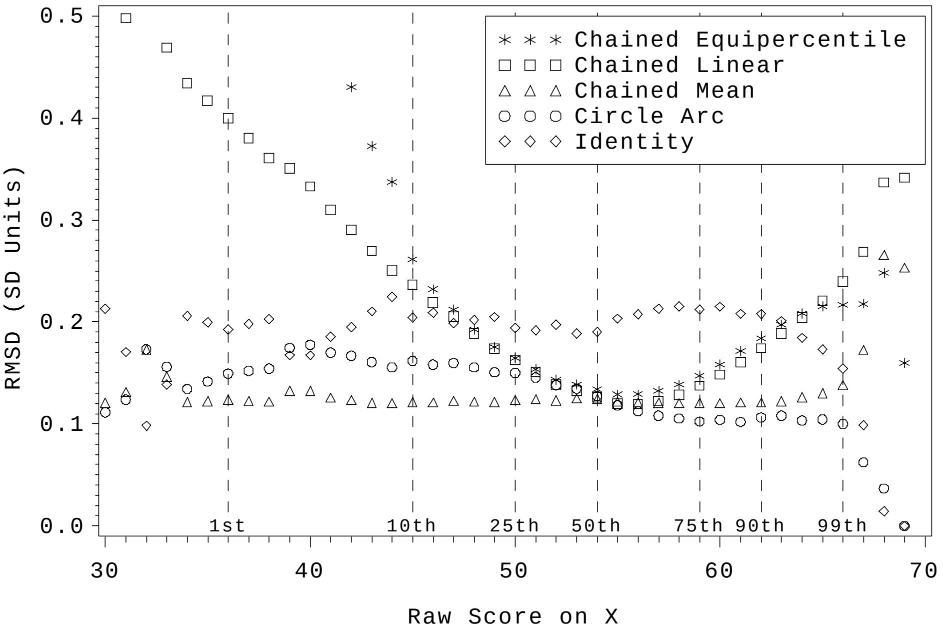


Figure A22. **Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 3A.**

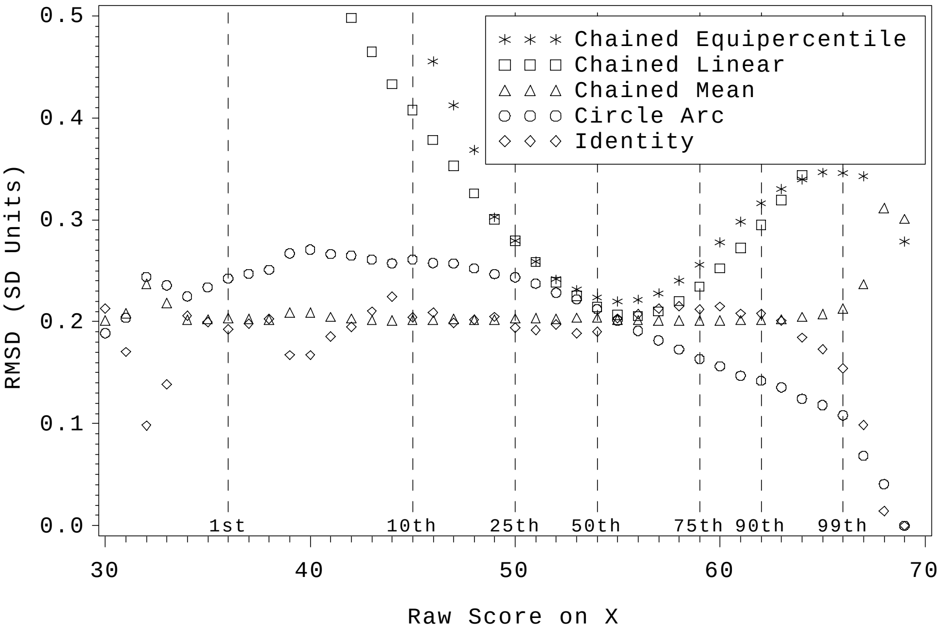


Figure A23. **Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 3A.**

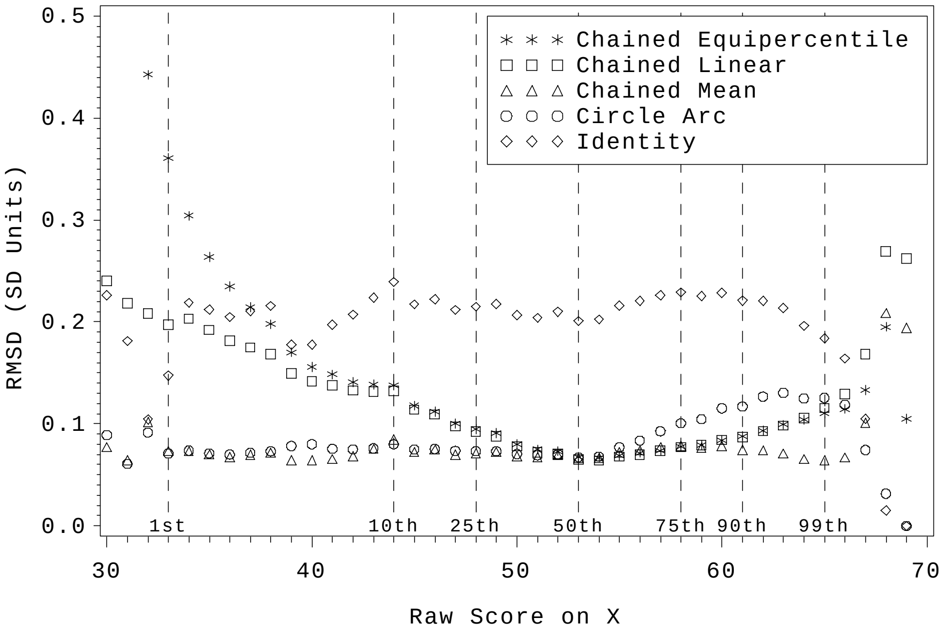


Figure A24. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 3B.**

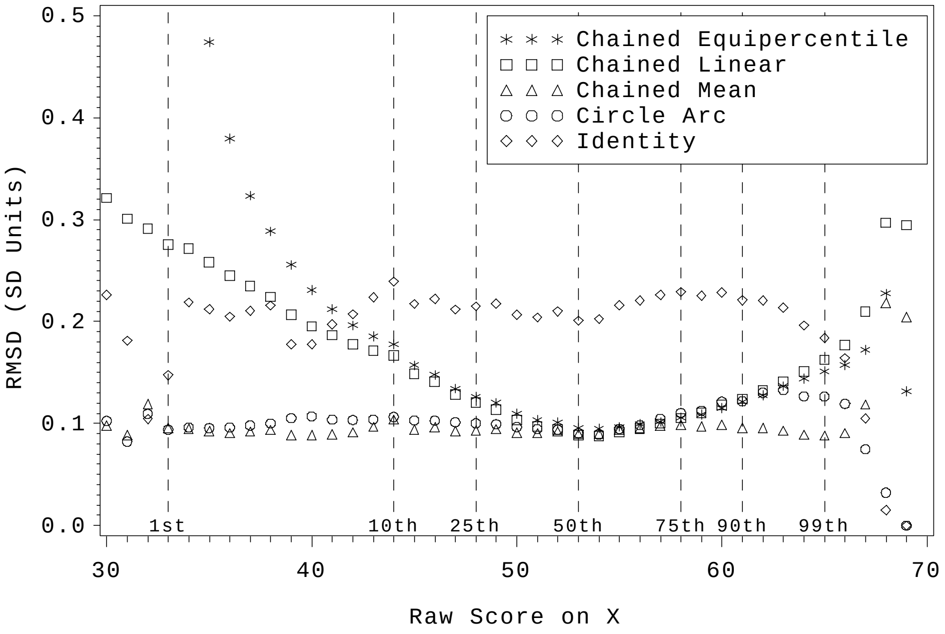


Figure A25. **Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 3B.**

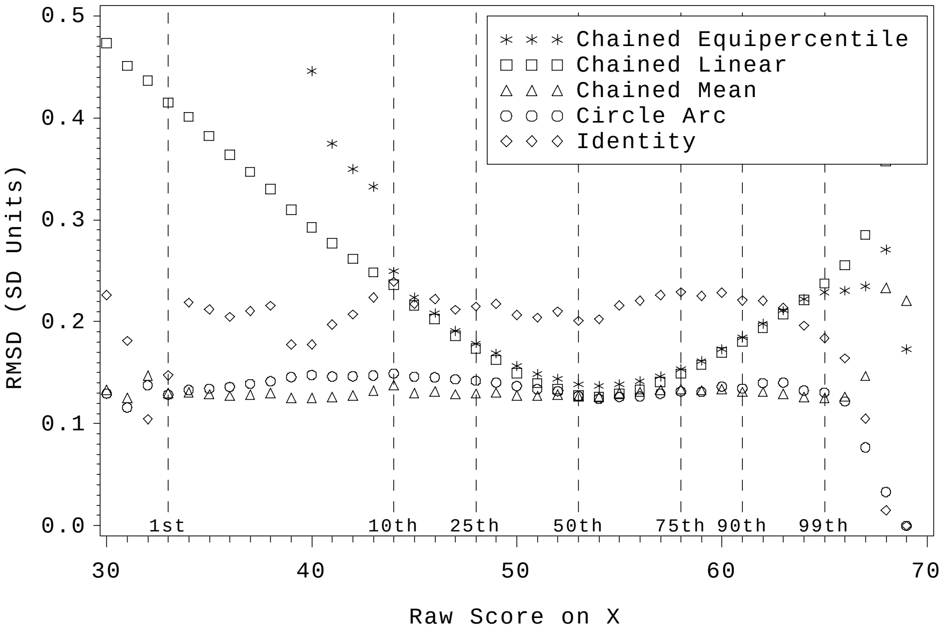


Figure A26. **Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 3B.**

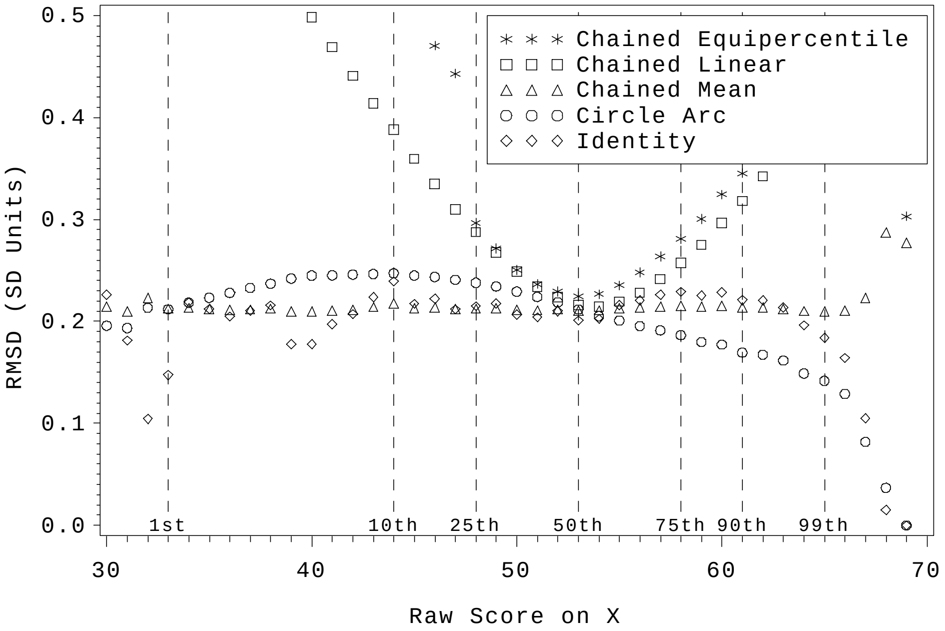


Figure A27. **Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 3B.**

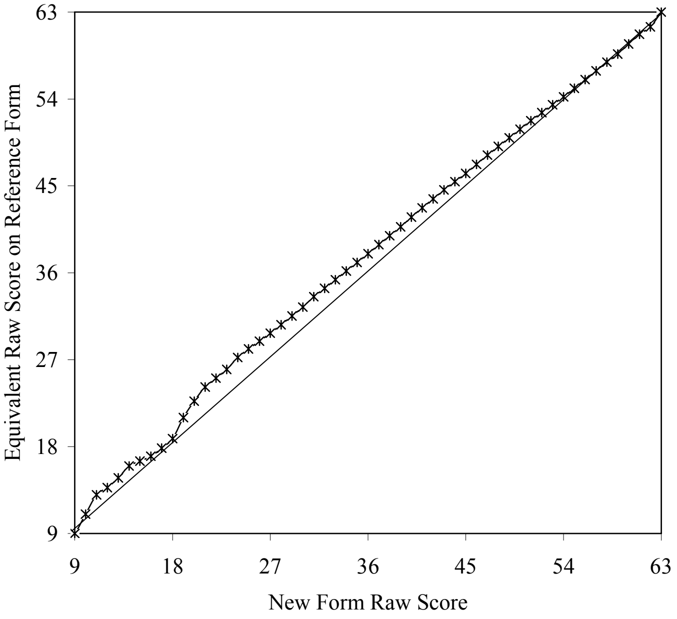


Figure A28. **Criterion equating for Research Form 4.**

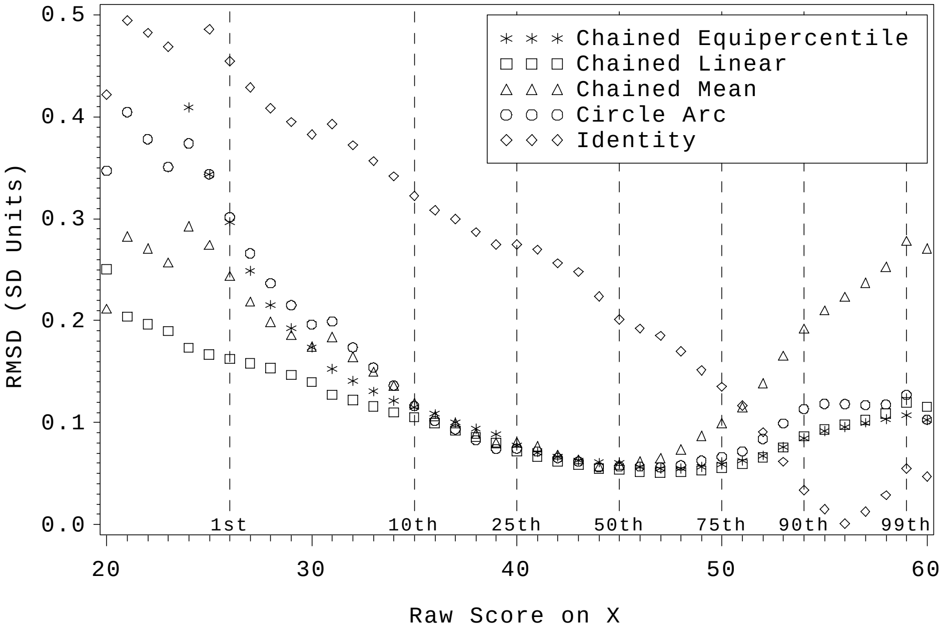


Figure A29. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 4A.**

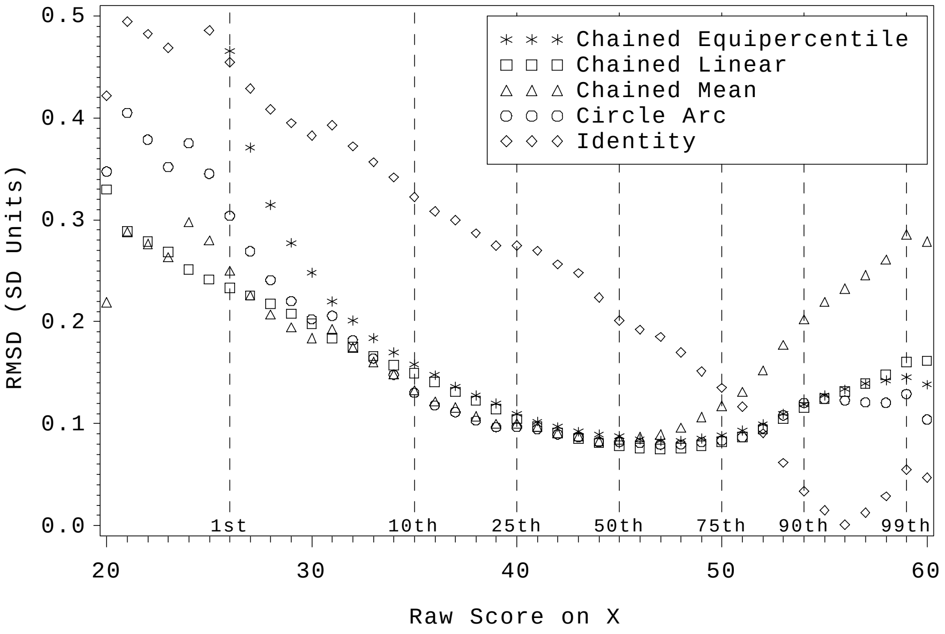


Figure A30. **Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 4A.**

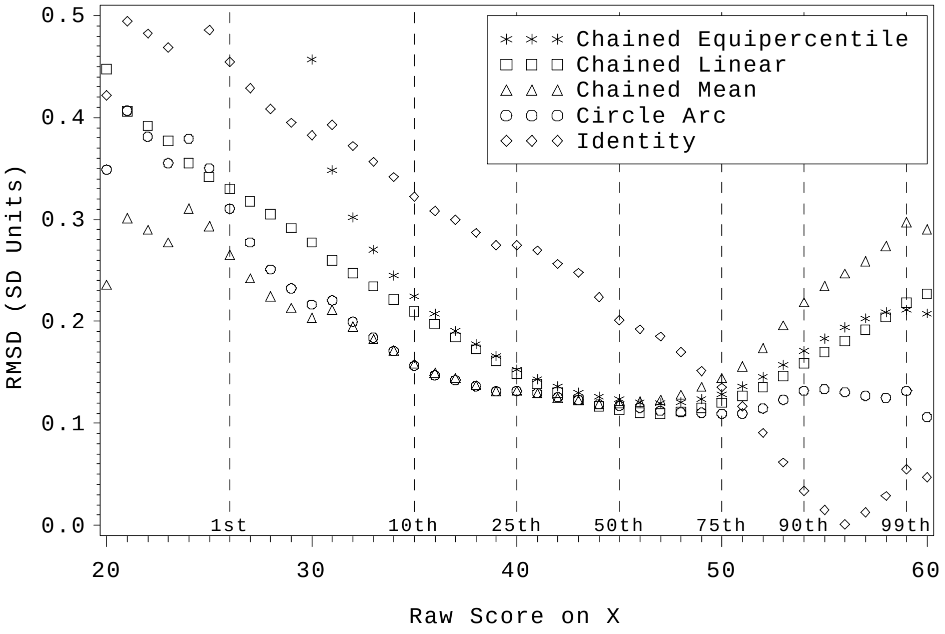


Figure A31. **Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 4A.**

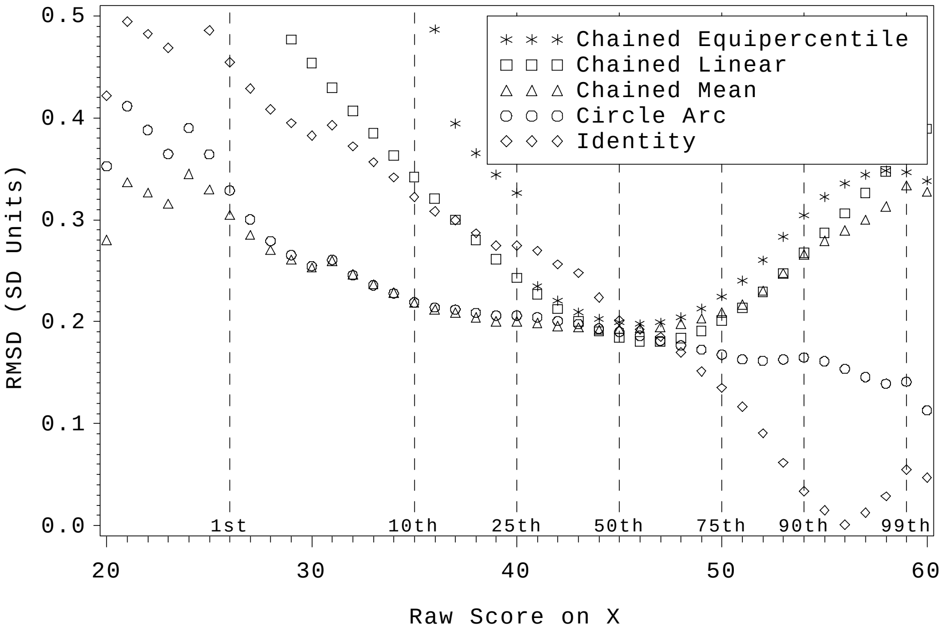


Figure A32. **Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 4A.**

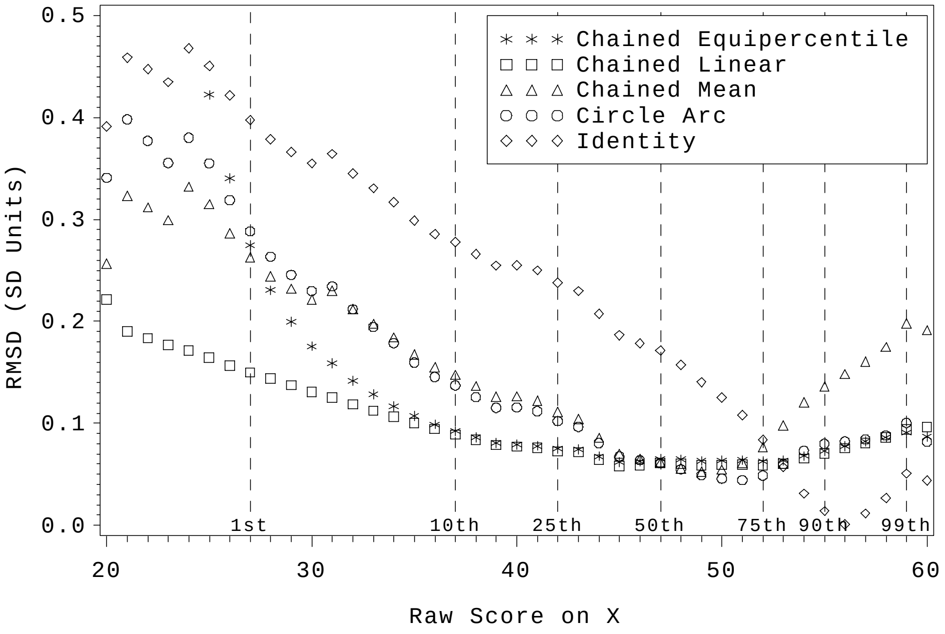


Figure A33. **Conditional root mean squared difference (RMSD): samples of 100 new-form examinees and 300 reference-form examinees, Equating 4B.**

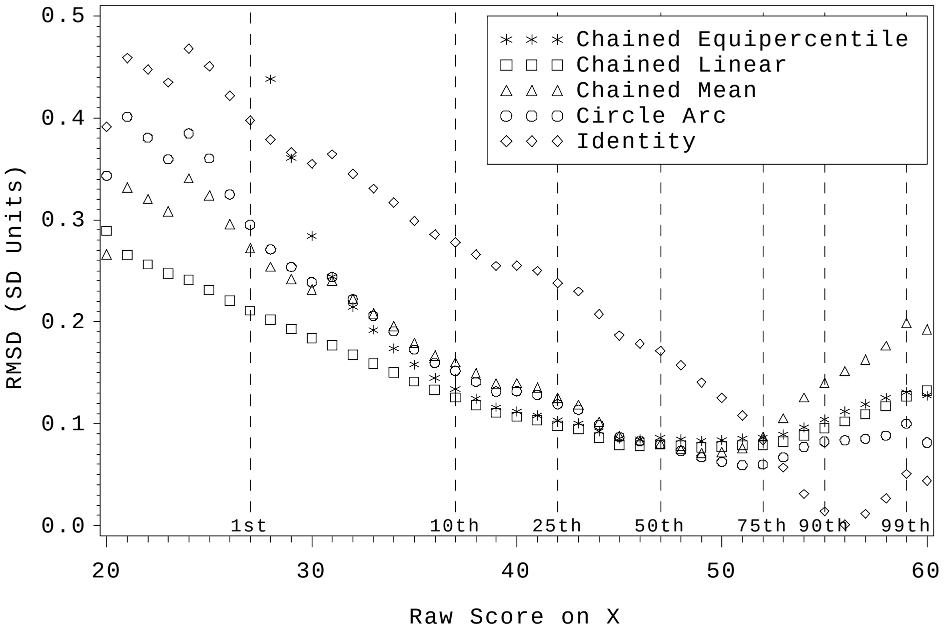


***Figure A34*. Conditional root mean squared difference (RMSD): samples of 50 new-form examinees and 150 reference-form examinees, Equating 4B.**

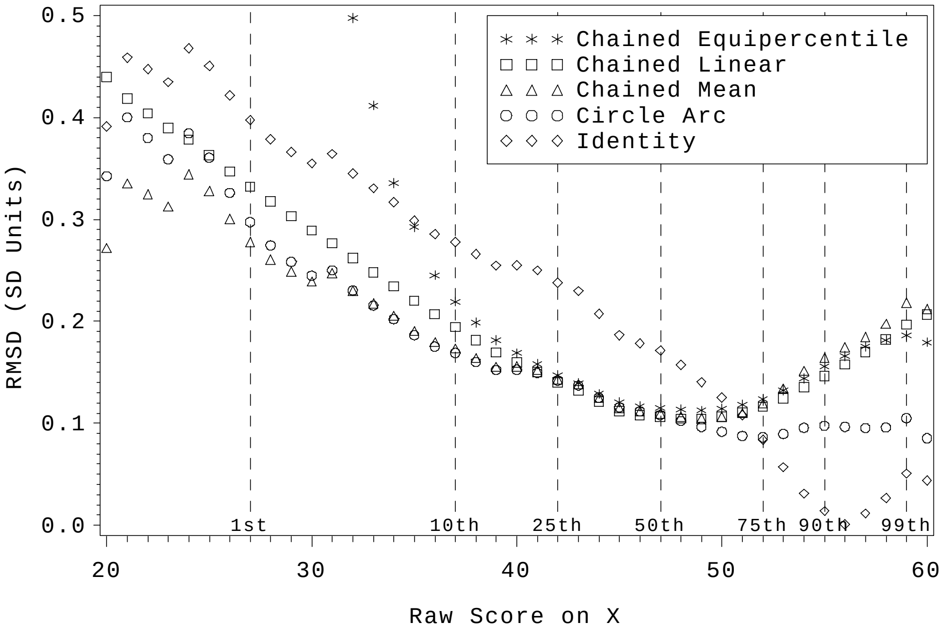


***Figure A35*. Conditional root mean squared difference (RMSD): samples of 25 new-form examinees and 75 reference-form examinees, Equating 4B.**

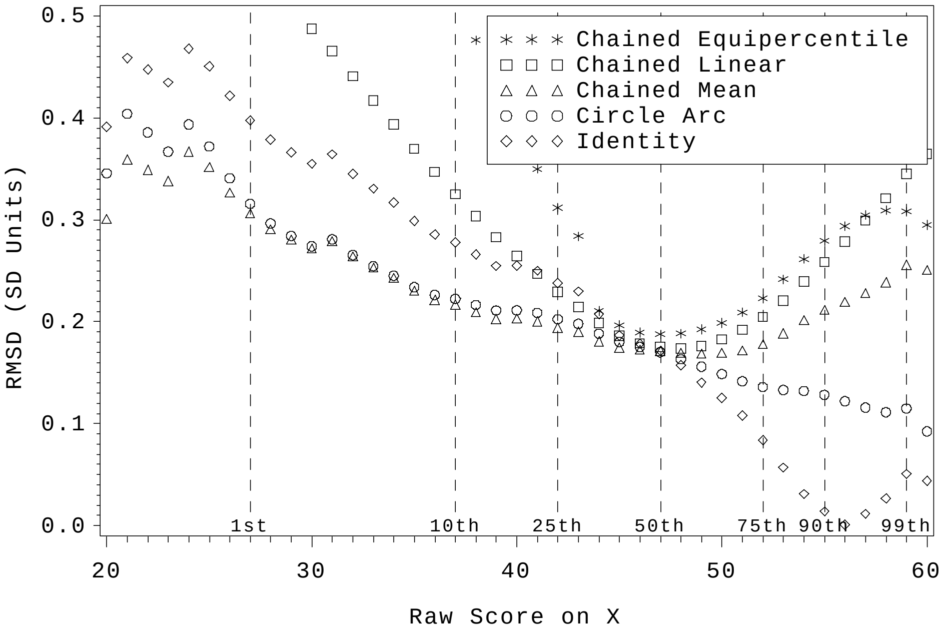


Figure A36**. Conditional root mean squared difference (RMSD): samples of 10 new-form examinees and 30 reference-form examinees, Equating 4B.**