Research Report

An Empirical Comparison of Methods for Equating With Randomly Equivalent Groups of 50 to 400 Test Takers

Samuel A. Livingston

Sooyeon Kim

February 2010

ETS RR-10-05

An Empirical Comparison of Methods for Equating With Randomly Equivalent Groups of 50 to 400 Test Takers

Samuel A. Livingston and Sooyeon Kim

ETS, Princeton, New Jersey

February 2010

Technical Review Editor: Dan Eignor

Technical Reviewers: Jim Carlson and Tim Moses

Abstract

A series of resampling studies investigated the accuracy of equating by four different methods in a random groups equating design with samples of 400, 200, 100, and 50 test takers taking each form. Six pairs of forms were constructed. Each pair was constructed by assigning items from an existing test taken by 9,000 or more test takers. The criterion equating was the direct equipercentile equating in the full group. Accuracy was described in terms of the root-mean-squared deviation (over 1,000 replications) of the sample equatings from the criterion equating. The equating methods investigated were equipercentile equating of smoothed distributions, linear equating, mean equating, and circle-arc equating; they were compared with each other and with the identity. Circle-arc equating produced the most accurate results for all sample sizes investigated, particularly in the upper half of the score distribution.

Key words: equating, random groups, small samples, circle-arc equating, resampling study

What does it mean to say that scores on two forms of a test are accurately equated? Angoff (1971, p.563; 1984, p.86) stated that scores on two different forms of a test "... may be considered equivalent if their corresponding percentile ranks in any given group are equal." Braun and Holland (1982, p. 15) and Holland and Dorans (2006, p. 202) used essentially the same definition, defining equivalence with reference to a particular population of test takers. The research we report in this paper is based on that definition of equating. We will consider an equating based on samples of test takers as accurate to the extent that it matches the equipercentile equating in the population from which those samples were drawn.

Equating the scores on two different forms of a test requires data collected in a way that connects the scores on the two forms. One such data collection plan is the *random groups* or *equivalent groups* equating design. It consists of administering the two forms of the test to nonoverlapping samples of the test-taker population, selecting those samples to be equal (as nearly as possible) in the abilities measured by the test.

The research reported here compares four methods for estimating an equating transformation from data collected according to such a plan. In this study, the samples were independent, non-overlapping, random samples of equal size. The sample size varied from 50 to 400 test takers—much smaller than the samples commonly used for equating in a random groups design. The four equating methods investigated were

1. equipercentile equating of smoothed distributions;
2. linear equating (i.e., setting means and standard deviations equal);
3. mean equating (i.e., adding or subtracting a constant); and
4. circle-arc equating.

The fourth method listed above—circle-arc equating—is new. Like mean equating, it estimates the entire equating transformation from a single empirically determined point on the equating curve. The curve is constrained to pass through that single empirically determined point and two end-points specified without reference to the data. The empirically determined point is the intersection of the mean scores on the test forms to be equated. The upper end-point is the intersection of the maximum possible scores; the lower end-point of the curve is the intersection of the lowest meaningful scores. The equating curve is estimated by decomposing it into a linear component and a curvilinear component. The linear component is the line connecting the end-points; the curvilinear component is a circle arc. (For a more thorough description of this method, including the computation formulas, see Livingston & Kim, 2008, 2009.)

The equating methods we compared vary systematically in the extent to which they substitute assumptions for data. The equipercentile equating method we used was an equating of the score distributions produced by the log-linear smoothing method of Holland and Thayer (1987), which requires the user to specify the features of the population distributions to estimate from the data. We chose to estimate only the mean, standard deviation, and skewness. Linear equating assumes that the population distributions to be equated differ only in their means and standard deviations. It requires only the population means and standard deviations to be estimated from the data. Mean equating assumes that the population distributions to be equated differ only in their means; it requires only the population means to be estimated from the data. Circle-arc equating also requires only the population means to be estimated from the data, although its assumptions cannot be stated simply in terms of the score distributions in the population.

Some authors have recommended using the identity as the equating transformation whenever the available samples of test takers are smaller than a prespecified size (Kolen & Brennan, 2004, pp. 289–290; Skaggs, 2005, p. 309). Although the identity is not really an equating method, we have included it in our comparisons. Using the identity as a substitute for equating makes the strongest assumption of all: that the population distributions of the scores to be equated do not differ. It does not require any features of the population distributions to be estimated from the data.

Our search of the literature revealed only two previous studies that investigated the accuracy of equating with small samples in a random groups design. Hanson, Zeng, and Colton (1994) compared linear equating and equipercentile equatings using several methods of smoothing (including no smoothing at all) in samples ranging in size from 100 to 3,000, using data from five different tests. No single method was the most accurate for all five tests. For four of the five tests investigated, log-linear smoothing that preserved only three moments of the observed distributions produced more accurate equating results than smoothings that preserved four or more moments.

Skaggs (2005) compared mean equating, linear equating, and equipercentile equatings using various degrees of pre-smoothing (including none at all) in samples ranging in size from 25 to 200, using data from only one test. Mean equating was the most accurate of the small-sample methods for below-average scores, but the least accurate for above-average scores. Linear equating was more accurate than equipercentile equating for below-average and near-average scores, but less accurate for scores more than one standard deviation above the mean.

Like the studies of Hanson et al. (1994) and Skaggs (2005), the studies we report here were resampling studies using real data, not simulated data. Each study consisted of drawing repeated samples from a population of test takers, applying the equating procedures, and comparing the sample results with a criterion equating.

Method

To evaluate the accuracy of an equating based on samples of test takers, it is necessary to know the population equating that the sample equating is intended to estimate. We began with six existing operational tests. Each of the six tests contained at least 98 items and had been taken by 9,000 or more test takers. Using each of these six tests as an item pool, we created two nonoverlapping research forms, equal in length (half as long as the operational form or slightly shorter) and parallel in content, but unequal in difficulty.

The test takers who had taken the operational test served as the population for the resampling studies. We computed the score of each test taker on each of the two research forms. We designated one of the two research forms as the new form and the other as the reference form. Using the scores of all the test takers, we performed a direct equipercentile equating of scores on the new form to scores on the reference form. This equating was the criterion equating for the resampling studies.

The basic procedure for each resampling study was as follows:

1. Specify the sample size to be investigated (the same size for each of the forms to be equated).

2. Perform 1,000 replications of the following procedure:

    a. From the test-taker population, draw a simple random sample of the required number of test takers. Use an odd-even split to divide this sample into two samples of the specified sample size. (Because in simple random sampling each individual is selected independently of the others, this procedure yields two independent, nonoverlapping random samples.) Designate the first sample as the new-form sample and the second as the reference-form sample.

   b. Equate the new form to the reference form in those samples by each of the equating methods to be compared.
   c. At each new-form raw-score level, for each equating method, compute the difference between the sample equating and the criterion equating.
3. At each new-form raw-score level, for each equating method, compute the root mean square average of the 1,000 differences computed in Step 2c above. This quantity is the root mean squared deviation (RMSD) of the sample equating results from the population equating. Also compute the RMSD for the identity, which is simply the difference between the identity and the population equating.

Thus, each resampling study involved a particular pair of test forms to be equated and a specified sample size (the same for both samples of test takers). Our investigation included six pairs of test forms and four specified sample sizes, for a total of 24 resampling studies. The appendix to this report contains a set of five graphs for each of the six pairs of test forms. The first graph in each set shows the criterion equating in comparison with the identity. Each of the remaining four graphs in the set shows the results of the resampling study at one of the specified sample sizes: 400, 200, 100, or 50 test takers for each form. The graph contains five curves, one for each small-sample equating method and one for the identity. The height of the curve at any given raw-score level shows the RMSD at that score level, expressed in standard-deviation units for comparability across the six pairs of test forms.

In the body of the report, the results of the resampling studies are summarized in four graphs, one graph for each specified sample size. For these graphs, the RMSD values were computed at specified percentiles of the score distribution and then averaged across the six different pairs of test forms. The six RMSD values that were averaged to determine each data point were computed at the same percentile of the score distribution and were expressed in standard-deviation units. Therefore, it is meaningful to average them across tests given to different populations.

The tests used as item pools for these studies were nationally administered teacher certification tests. Table 1 shows, for each test, the subject of the test, the number of items in the test, the number of test takers in the population, and the mean and standard deviation of their raw (number correct) scores. Table 2 shows a statistical comparison of the two research forms created from each of the tests in Table 1, based on the scores of the full population. On three of the six pairs of test forms, the mean scores of the population differed by slightly more than one fourth of a standard deviation. On two other pairs of test forms, the mean scores differed by one sixth or one seventh of a standard deviation. On one pair of test forms, the mean scores differed by only about one twelfth of a standard deviation. The last column of Table 2 shows the largest difference between cumulative distribution functions for the two test forms in the full population. (This statistic is sometimes called the Kolmogorov D-statistic.) Multiplied by 100, it is the largest difference between the two forms in the percentile rank of any given raw score. For example, on Test 1, there is at least one raw score for which the percentile rank on the new form is 11 points lower than the percentile rank for the same raw score on the reference form. On Test 3, there is no raw score for which the difference in percentile ranks is more than three percentile points.

Table 1

Tests Used as Item Pools

Test	Subject	Number of items	Number of test takers	Mean	SD
1	Social studies: content knowledge	130	9,240	76.38	15.29
2	Elementary education: content knowledge	120	15,525	80.62	14.41
3	English language, literature, and composition: content knowledge	120	15,401	90.61	14.28
4	Speech-language pathology	150	13,054	93.03	13.92
5	Educational leadership: administration and supervision	117	9,597	82.56	12.30
6	Fundamental subjects: content knowledge	98	14,407	73.20	11.32

Table 2

Test Forms Created From Each Item Pool

Test	Number of items in each form	Form *X*			Form *Y*			Standardized mean difference	Largest difference in cdf
		M	SD	Skew	M	SD	Skew		
1	62	38.49	7.64	-0.04	36.43	7.93	0.04	0.27	-.11
2	60	40.96	7.43	-0.37	39.66	7.60	-0.33	0.17	-.07
3	59	45.35	7.20	-0.60	44.77	7.43	-0.62	0.08	-.03
4	74	45.82	7.30	-0.35	46.88	7.40	-0.49	-0.14	.07
5	58	40.31	6.42	-0.55	42.00	6.52	-0.60	-0.26	.13
6	47	34.44	6.06	-0.49	36.20	5.47	-0.60	-0.30	.12

Note. cdf = cumulative distribution function.

Results

The findings of our resampling studies are shown in a series of four graphs. Each graph shows the RMSD curves for one sample size, computed at nine selected percentiles of the score distribution and averaged over the six pairs of forms investigated. The RMSD values are expressed in standard deviation units for comparability across the six pairs of forms. The averaging process used a root-mean-square procedure, squaring the RMSD values, averaging them, and then taking the square root. The percentiles at which the RMSDs were computed were based on the distribution of scores on the new form in the full test taker population for that test. The vertical scale in the graphs extends from 0.0 to 0.3 standard deviation (*SD*) units; RMSD values of 0.1 and 0.2 *SD* are indicated by dotted horizontal lines. In some cases, the RMSD curves at the lowest and highest percentiles are outside this range. We chose to focus on this range because the RMSD values for the identity were all within this range. We considered it more important to show a clear comparison of RMSD values within this range than to show how far beyond this range the RMSD values for the least accurate methods were.

Figure 1 shows the RMSD curves for equating in samples of 400 test takers for each form. The most obvious result is that with samples of this size, all the equating methods produced much more accurate results than using the identity—hardly a surprise, since the test forms differed in difficulty. At the 25th and 50th percentiles of the new-form score distribution, all the equating methods yielded equally accurate results. Below the 25th percentile, the two methods that required estimation of only the population means (mean equating and circle-arc equating) outperformed the two methods that required the estimation of more than one parameter of each score distribution (linear equating and equipercentile equating with three-moment smoothing). Above the 50th percentile, mean equating produced less accurate results than linear equating, equipercentile equating produced slightly more accurate results than linear equating, and circle-arc equating produced the most accurate results.

Figure 2 shows the RMSD curves for equating in samples of 200 test takers for each form. The RMSD values are larger than those for the samples of 400 test takers, particularly in the middle of the score distribution, but the comparisons between equating methods are generally similar. The main difference is that, for above-average scores, mean equating performed nearly as well as linear equating. Also, at the 90th percentile and above, the advantage of circle-arc equating over linear equating and equipercentile equating was greater with samples of 200 than with samples of 400.

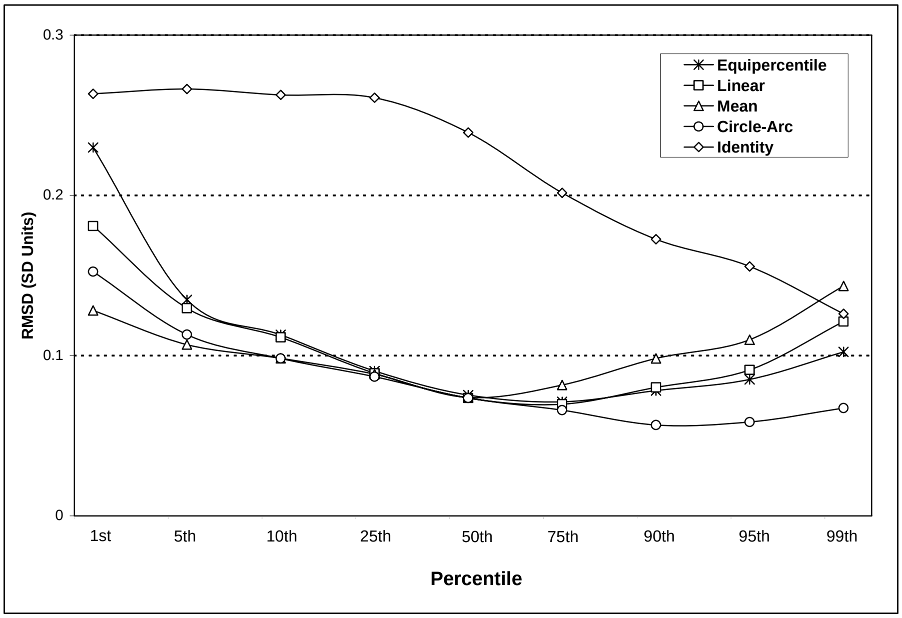

***Figure 1*. Root mean squared deviation (RMSD) of small-sample equating methods with samples of 400 test takers.**

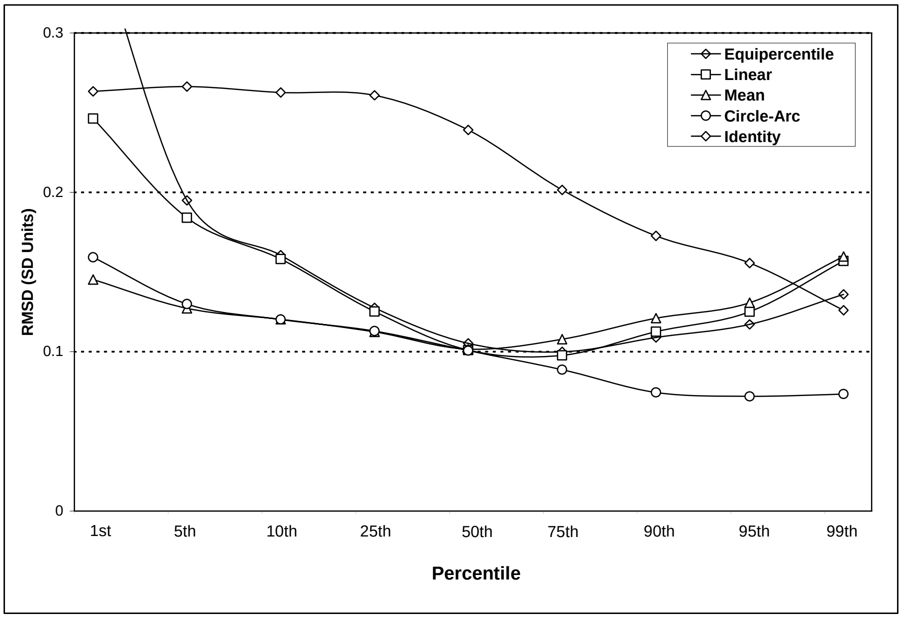

***Figure 2*. Root mean squared deviation (RMSD) of small-sample equating methods with samples of 200 test takers.**

Figure 3 shows the RMSD curves for equating in samples of 100 test takers. The RMSD values are large enough to make the equating at the 95th and 99th percentiles less accurate than the identity for all the small-sample equating methods except the circle-arc method. The comparisons among the four methods are similar to those for the larger samples, although the rank ordering of the methods at the 95th and 99th percentiles is different. Also, the difference in accuracy between the circle-arc method and the three other methods was greater with the samples of 100 than with the larger samples.

Figure 4 shows the RMSD curves for equating in samples of 50 test takers. With samples of this size, the RMSD values for all the equating methods and for the identity are 0.2 *SD* or larger throughout the lower half of the score distribution. At the 50th percentile, all four equating methods performed similarly. Below the 50th percentile, mean equating and circle-arc equating were much more accurate than the linear and equipercentile equating methods. Above the 50th percentile, the circle-arc method was much more accurate than any of the other equating methods. The circle-arc method was the only method that produced more accurate results than the identity throughout the score distribution.

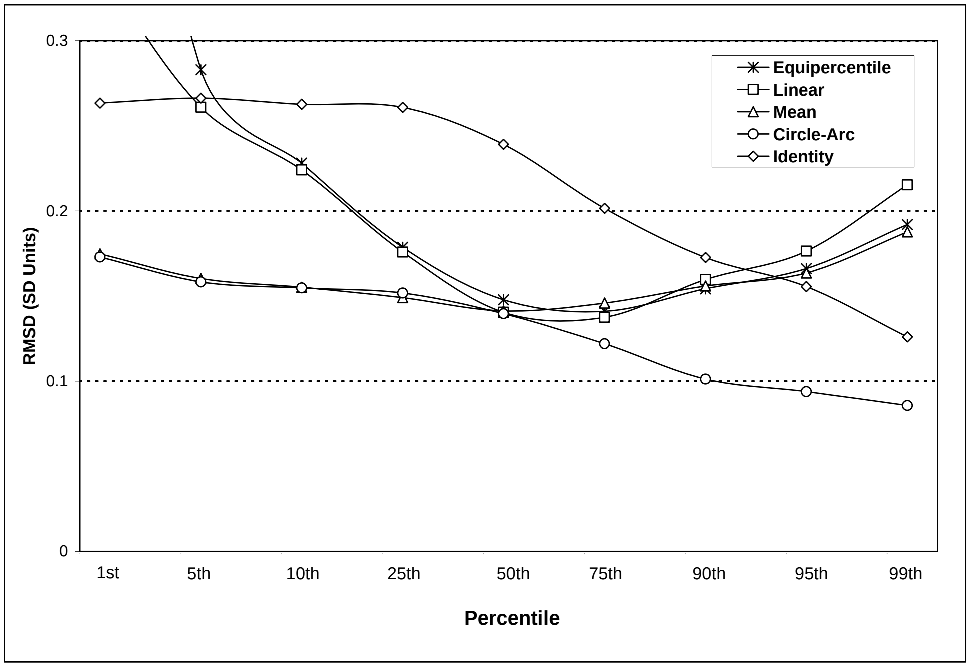

***Figure 3*. Root mean squared deviation (RMSD) of small-sample equating methods with samples of 100 test takers.**

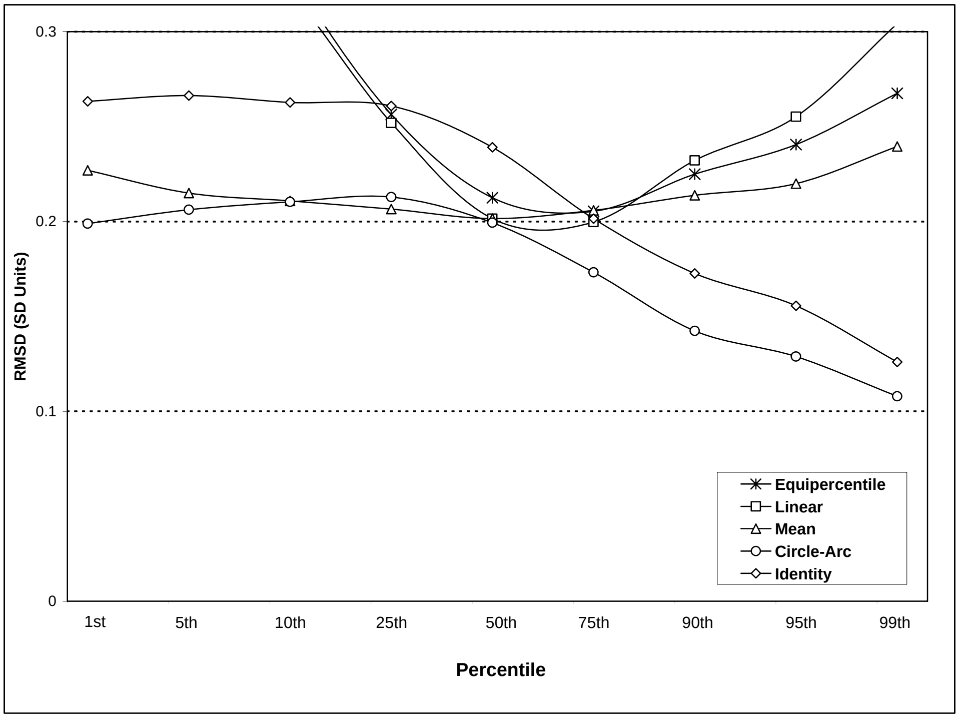

***Figure 4*. Root mean squared deviation (RMSD) of small-sample equating methods with samples of 50 test takers.**

Discussion

In this investigation of four methods for equating in a random groups design, the circle-arc method produced the most accurate results. The smaller the sample, the greater its advantage in accuracy over the other methods. It was particularly accurate for equating at high score levels. All four methods were about equally accurate at the 50th percentile. As the scores increased beyond the 75th percentile, the difference in accuracy between the circle-arc method and the three other methods became larger. For low scores— at the 25th percentile and below—both circle-arc equating and mean equating produced more accurate results than the other methods, particularly for samples of 200 or fewer test takers.

The success of the circle-arc method appears to be attributable to two features. First, it requires the estimation of only a single point on the equating curve, i.e., the intersection of the mean scores on the new form and reference form. Second, it equates the maximum possible score on the new form to the maximum possible score on the reference form by specifying the intersection of those two scores as the upper end-point of the equating curve. The first feature enables the method to work well with limited amounts of data. The second feature results in an estimated equating curve that resembles the curves typically produced by equipercentile equating.

Mean equating performed well for average and below-average scores, but not for high scores. This result is consistent with the results reported by Skaggs (2005). The main limitation of mean equating seems to be its inability to model a curvilinear equating relationship. When test forms differ in difficulty, their distributions in a population tend to be unequally skewed. The easier form tends to disperse the weaker test takers more widely through the lower and middle parts of the score range while bunching the stronger test takers more closely together in the higher parts of the score range. The harder form has the opposite effect. The difference in the shape of the distributions leads to a curvilinear equating relationship.

Linear equating and equipercentile equating both performed poorly for scores below the 25th percentile and for scores above the 90th percentile, especially with samples of 200 or fewer test takers. The problem with equipercentile equating is that the percentile ranks of scores in those regions are not accurately estimated in small samples. Linear equating appears to have two limitations: it cannot estimate a curvilinear relationship, and the slope of the conversion often is not estimated accurately when the samples are small.

The use of the identity would not have been a good substitute for equating the test forms constructed for these studies. In constructing those test forms, we deliberately made the new form and reference form unequal in difficulty; otherwise, no equating would be necessary. In practice, if test forms can be assembled from items for which highly accurate difficulty estimates are available, it may be possible to make the forms so nearly equal in difficulty at all ability levels that no equating is necessary. In that case, the use of the identity would be preferable to equating the scores on the basis of small-sample data. For the forms created for these studies, the use of the identity compared favorably with linear equating and equipercentile equating when the equating samples included only 50 test takers. But even with samples of that size in a random groups equating design, mean equating was more accurate than the identity for all percentiles below the 75th, and circle-arc equating was more accurate than the identity throughout the entire score range.

The circle-arc equating method used in this study is the method identified in Livingston & Kim (2008) as “Circle-Arc Method 2.” This method, applied to population data, is not truly an equating method, because it is not symmetric. There is another version of the circle-arc method, identified in Livingston & Kim (2008) as “Circle-Arc Method 1,” which is symmetric. Applied to the data in these studies, Method 1 produced results very close to those of Method 2. However, circle-arc equating is not intended as an alternative way to define the equating transformation. Its purpose is to estimate the equipercentile equating function in the population (which is a symmetric function) on the basis of data from small samples of test takers. How large do the samples in a random groups equating design have to be for equipercentile equating (with smoothing of the score distributions) to outperform circle-arc equating? To answer that question would require another set of studies like these, with larger samples. In these studies, with samples of 50 to 400 test takers, circle-arc equating provided a better estimate than could be obtained by equipercentile equating of smoothed score distributions, or by linear equating, or by mean equating, or by the use of the identity as a substitute for equating.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

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

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York, NY: Academic Press.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Rep. No. 94-4). Iowa City, IA: American College Testing, Inc.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.

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

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

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

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

Appendix

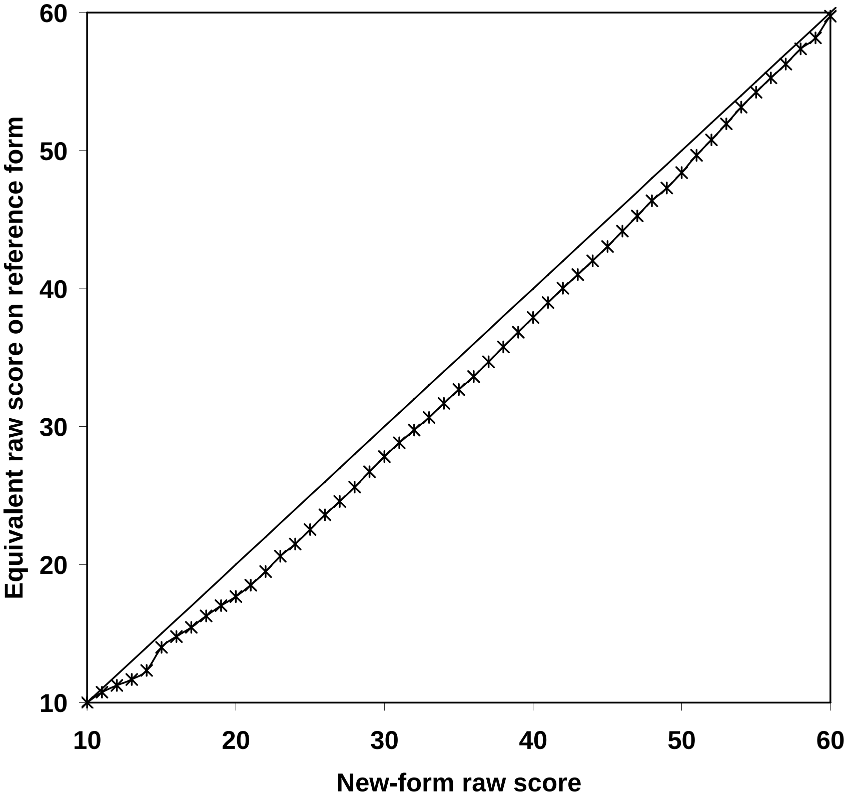

Figure A1. **Criterion equating for Test 1.**

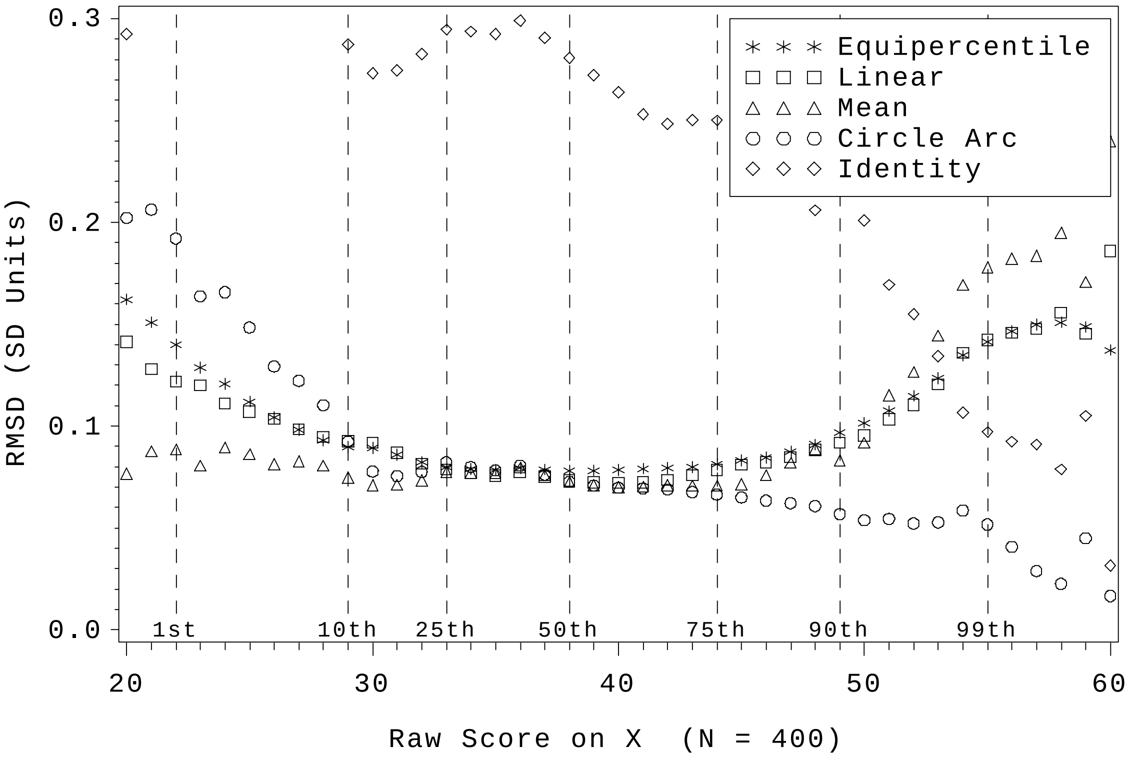

Figure A2. **Conditional root mean squared deviation: samples of 400, Test 1.**

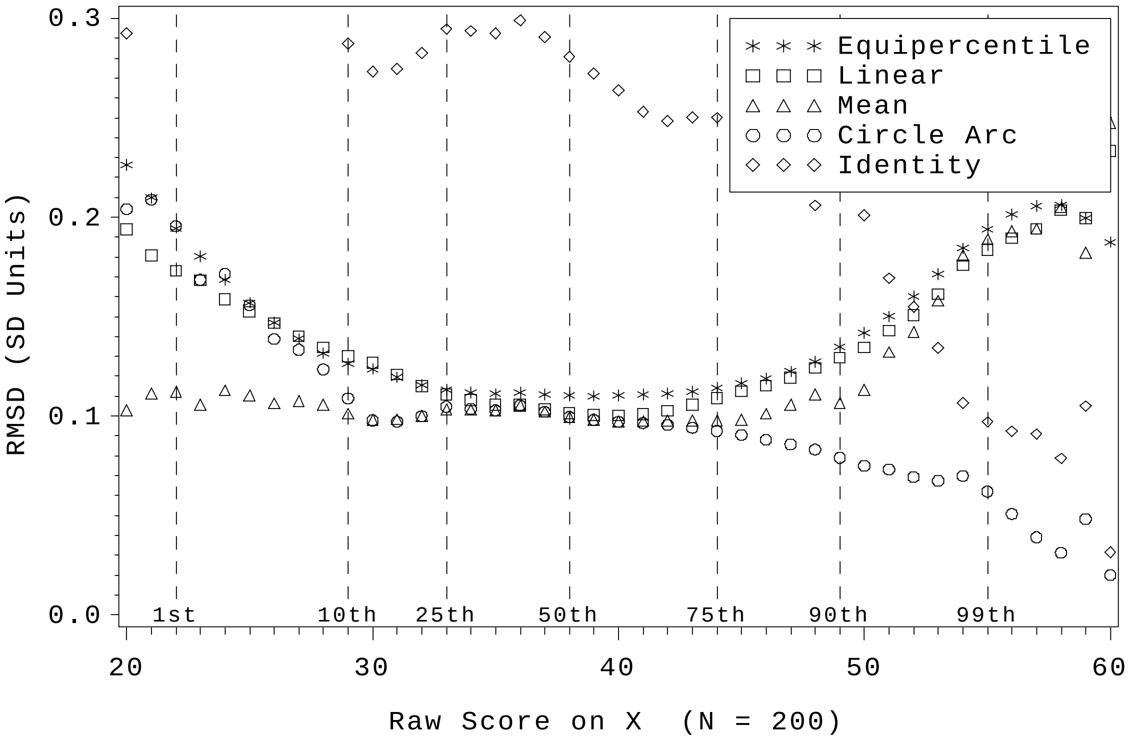

Figure A3. **Conditional root mean squared deviation: samples of 200, Test 1.**

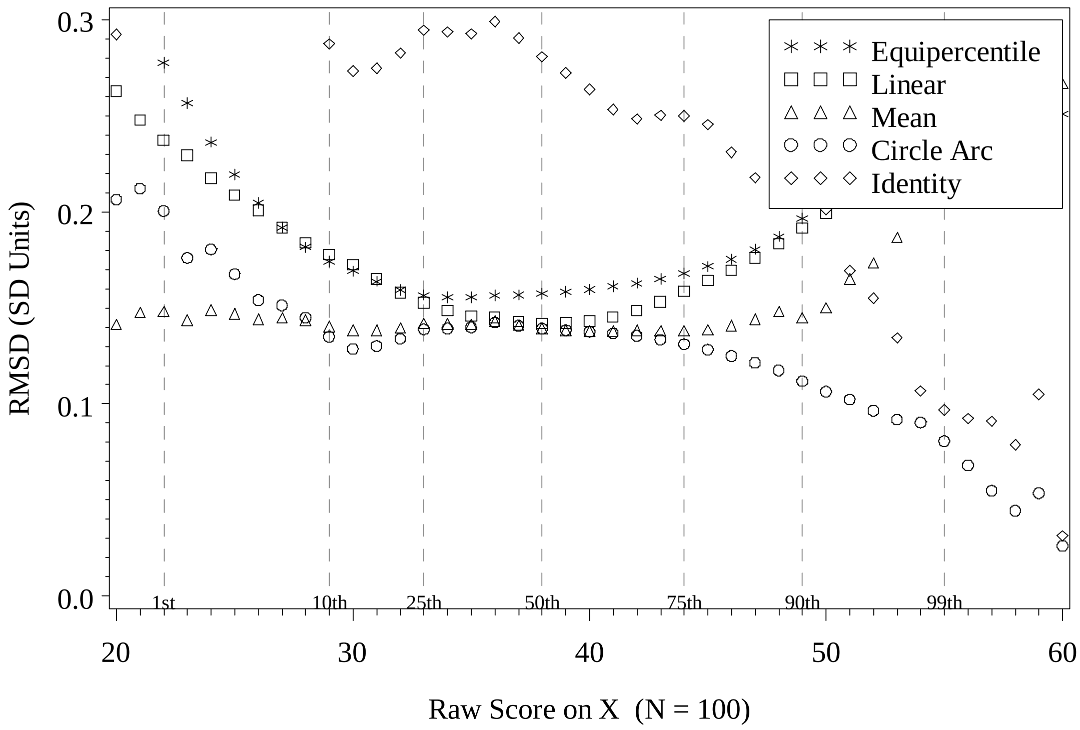

Figure A4. **Conditional root mean squared deviation: samples of 100, Test 1.**

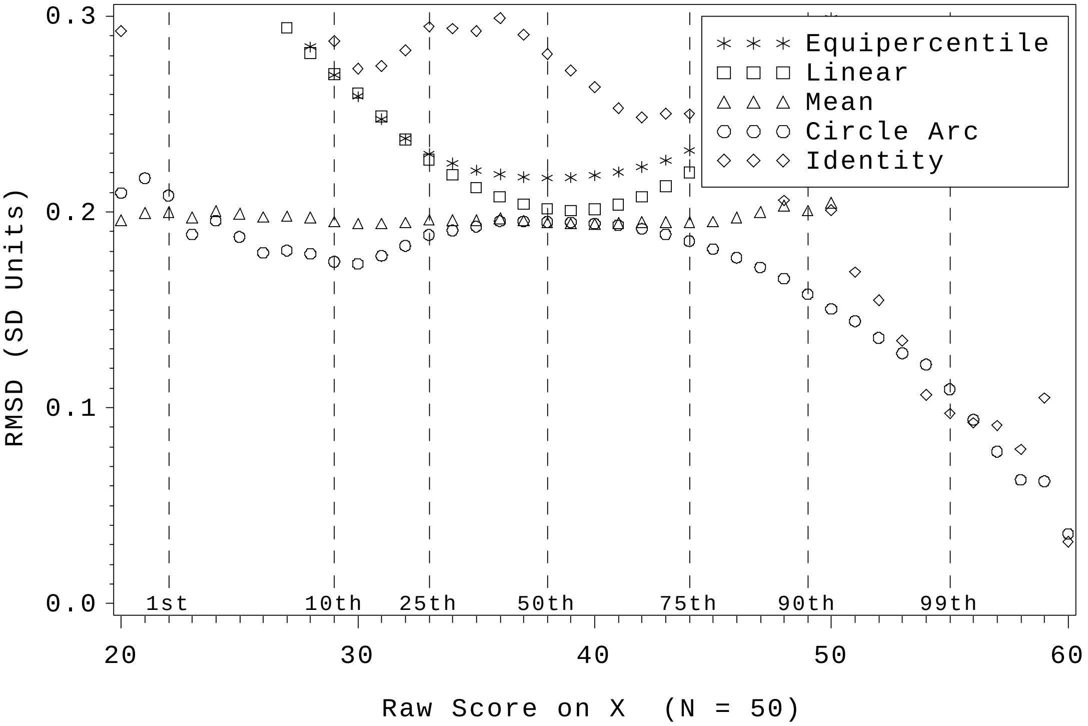

Figure A5. **Conditional root mean squared deviation: samples of 50, Test 1.**

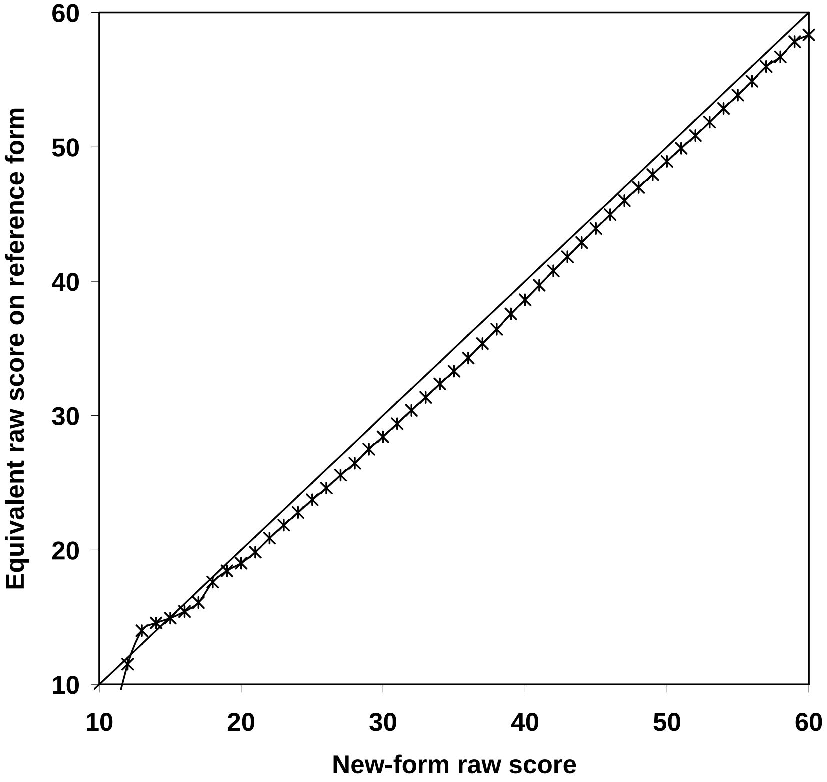

Figure A6. **Criterion equating for Test 2.**

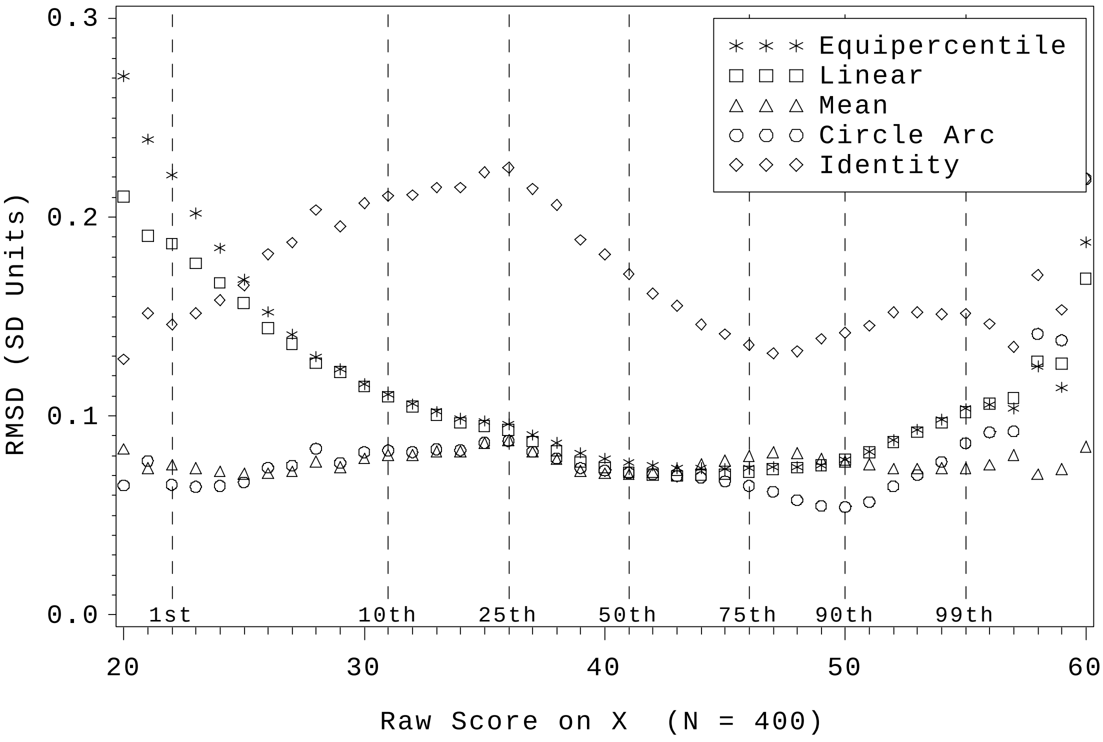

Figure A7. **Conditional root mean squared deviation: samples of 400, Test 2.**

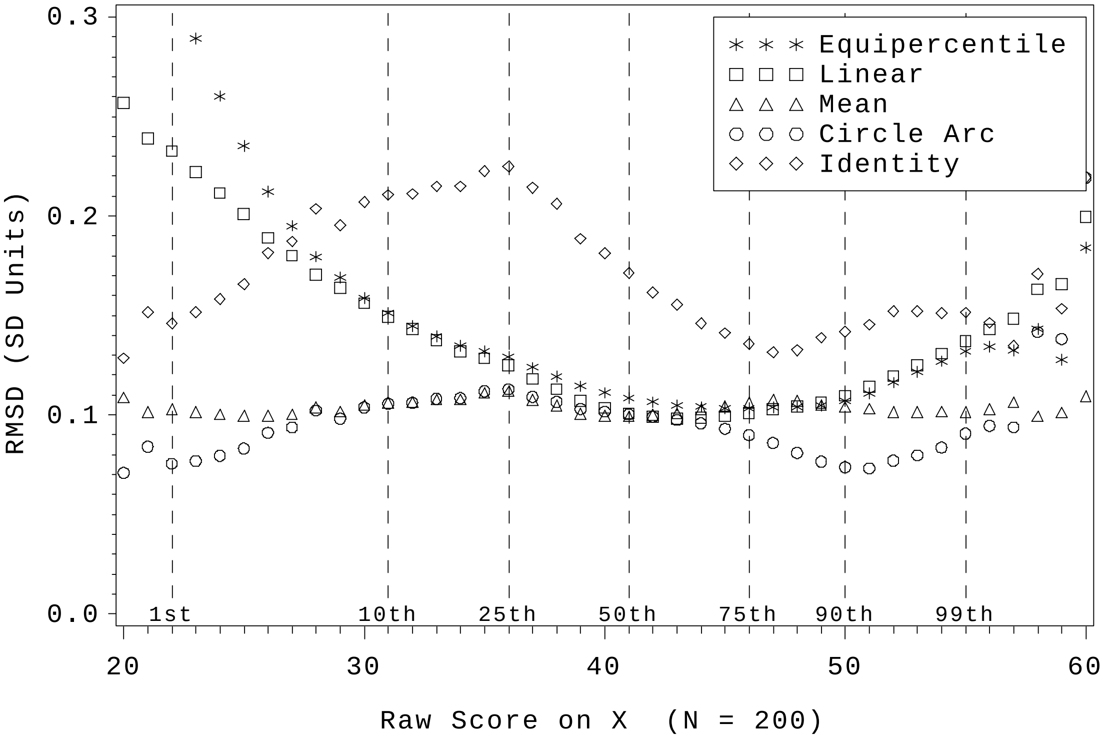

Figure A8. **Conditional root mean squared deviation: samples of 200, Test 2.**

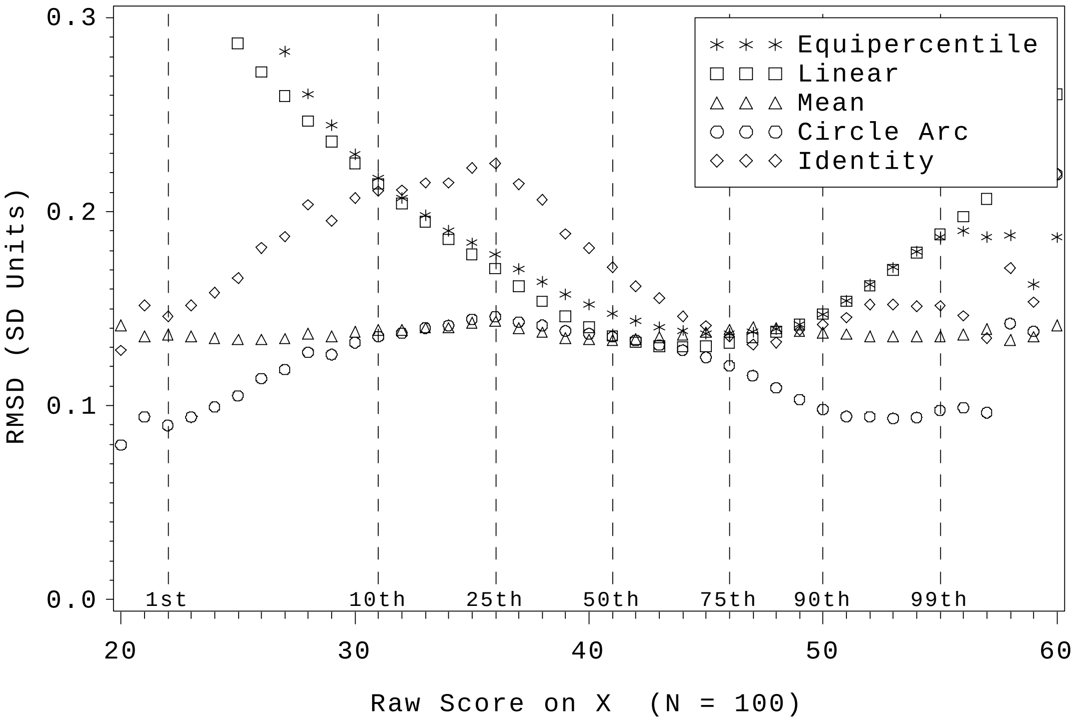

Figure A9. **Conditional root mean squared deviation: samples of 100, Test 2.**

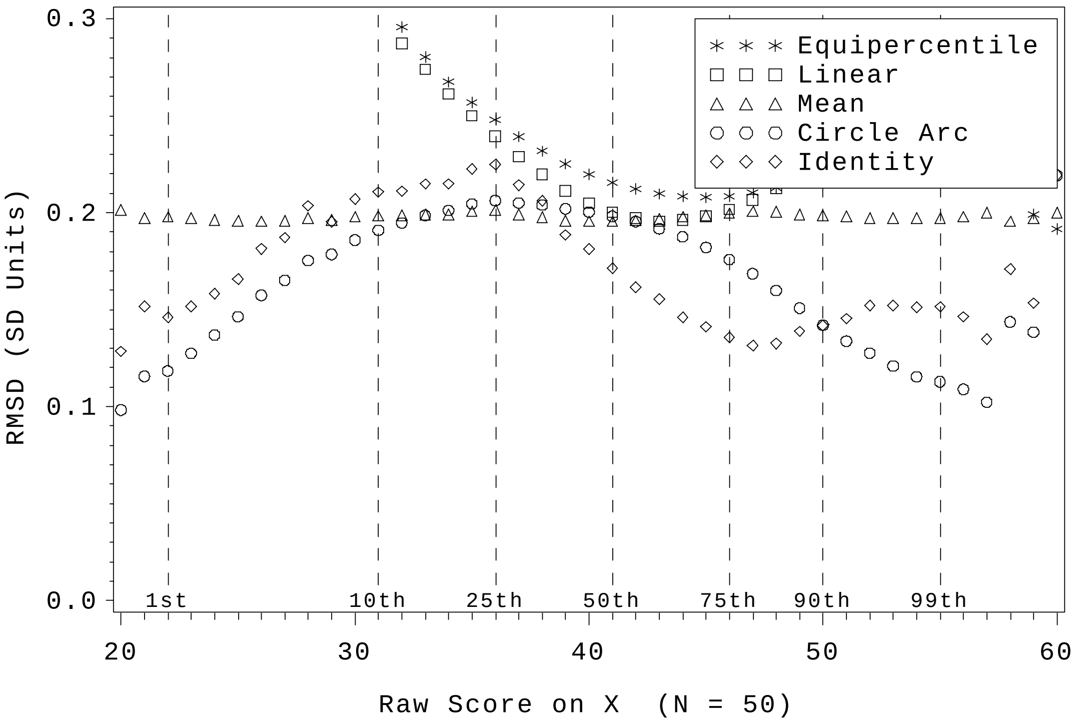

Figure A10. **Conditional root mean squared deviation: samples of 50, Test 2.**

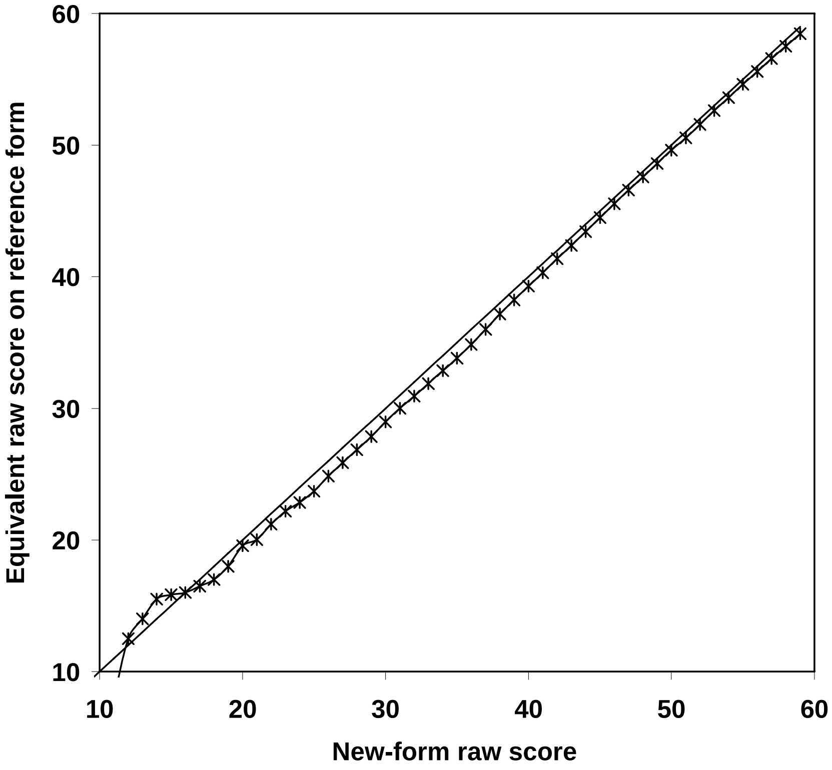

Figure A11. **Criterion equating for Test 3.**

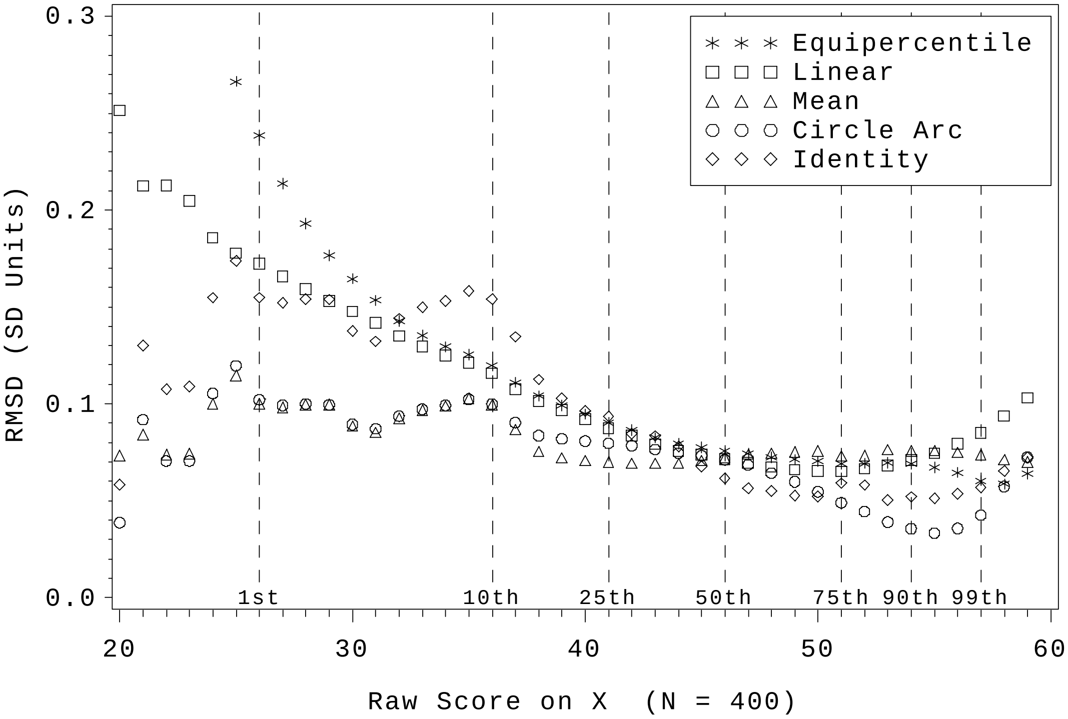

Figure A12. **Conditional root mean squared deviation: samples of 400, Test 3.**

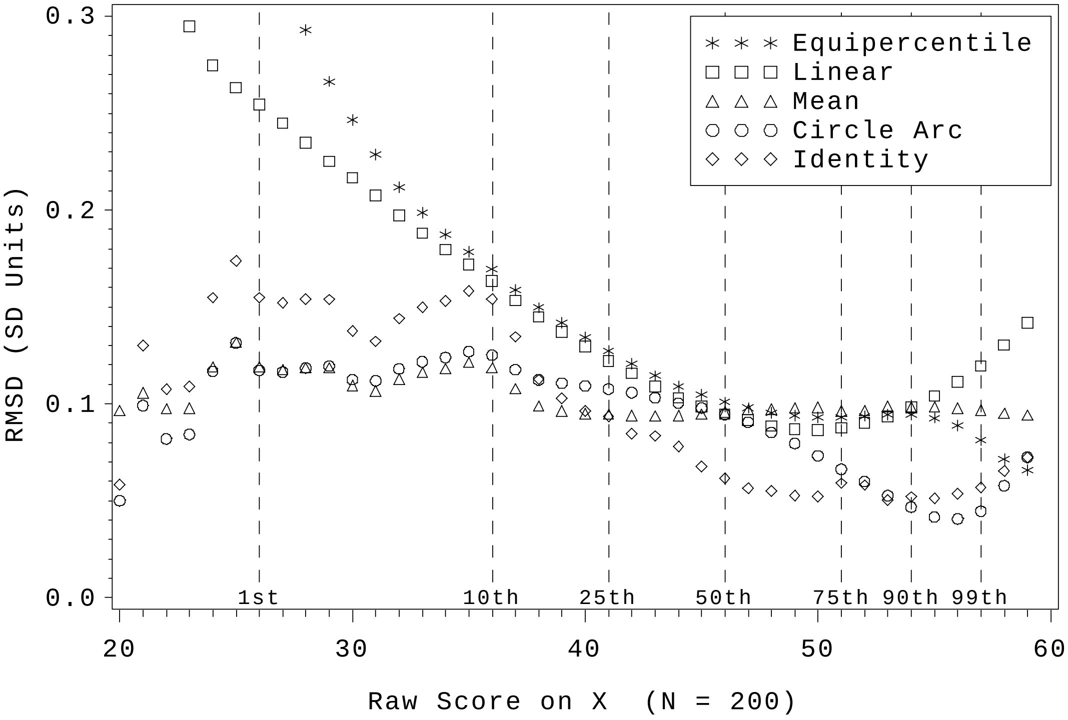

Figure A13. **Conditional root mean squared deviation: samples of 200, Test 3.**

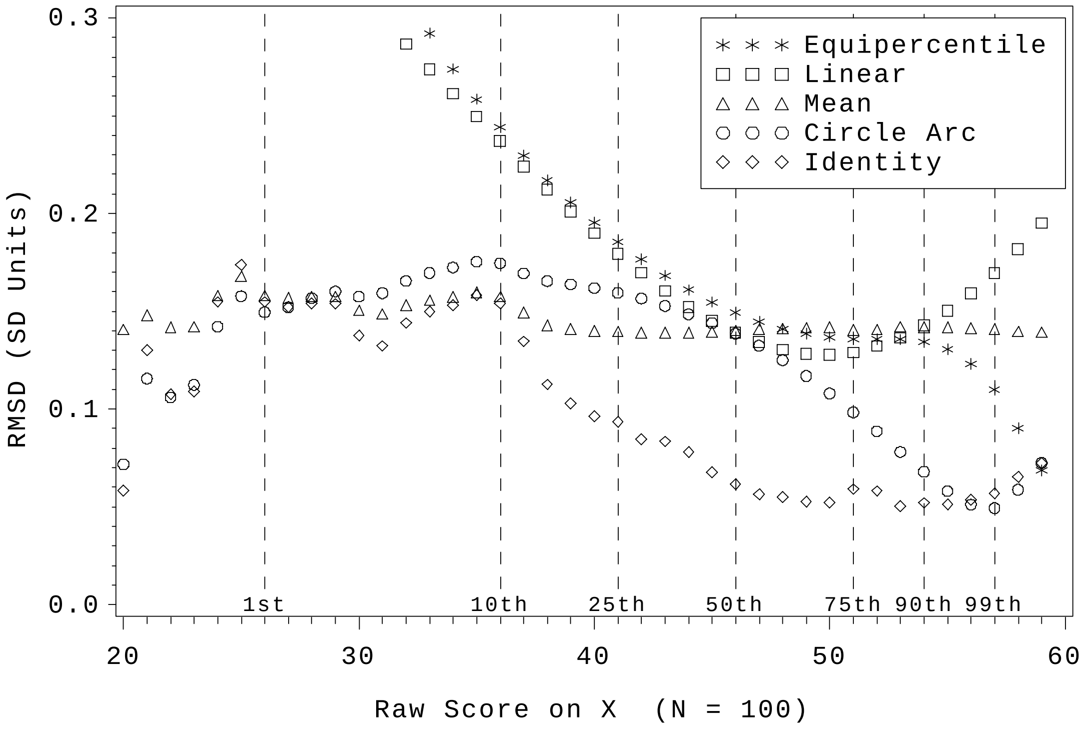

Figure A14. **Conditional root mean squared deviation: samples of 100, Test 3.**

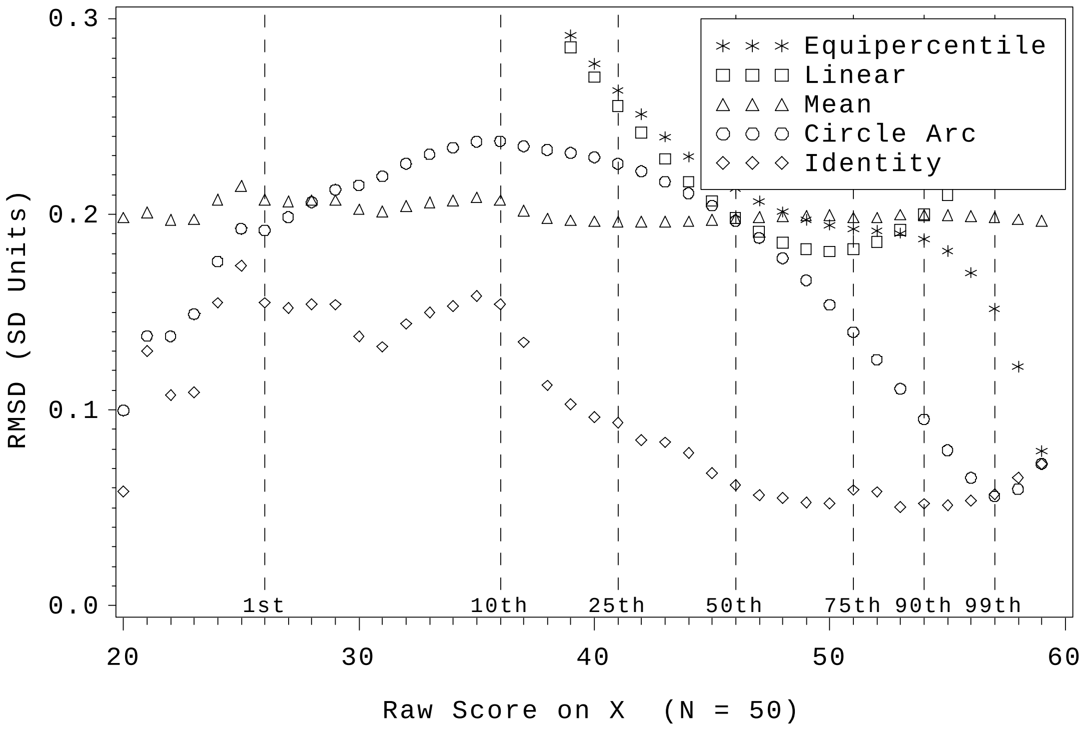

Figure A15. **Conditional root mean squared deviation: samples of 50, Test 3.**

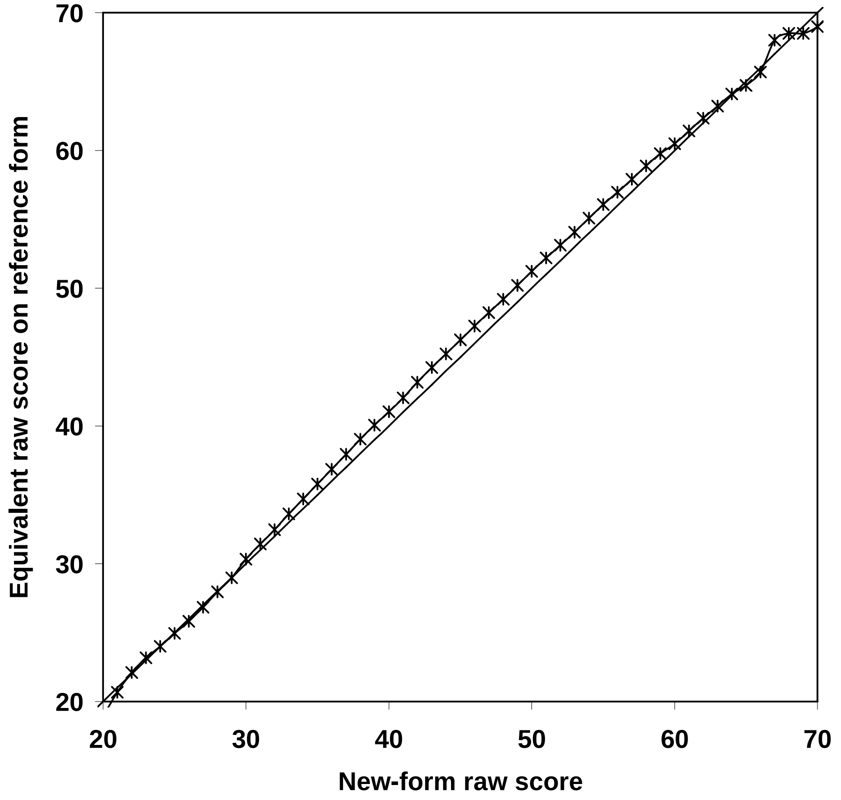

Figure A16. **Criterion equating for Test 4.**

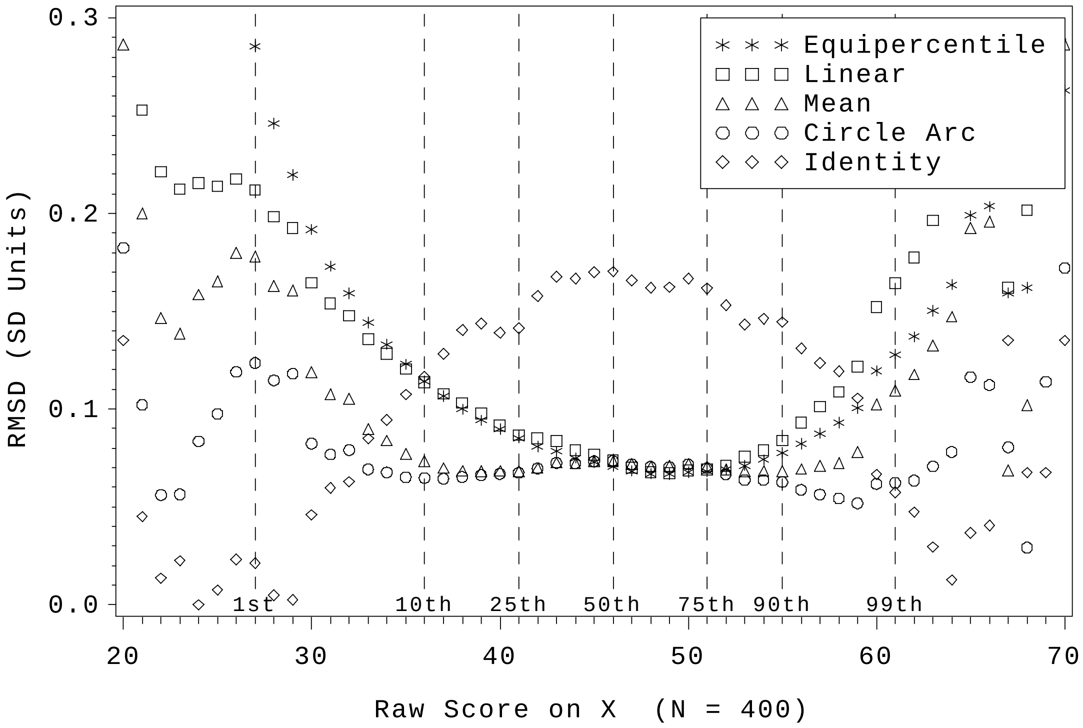

Figure A17. **Conditional root mean squared deviation: samples of 400, Test 4.**

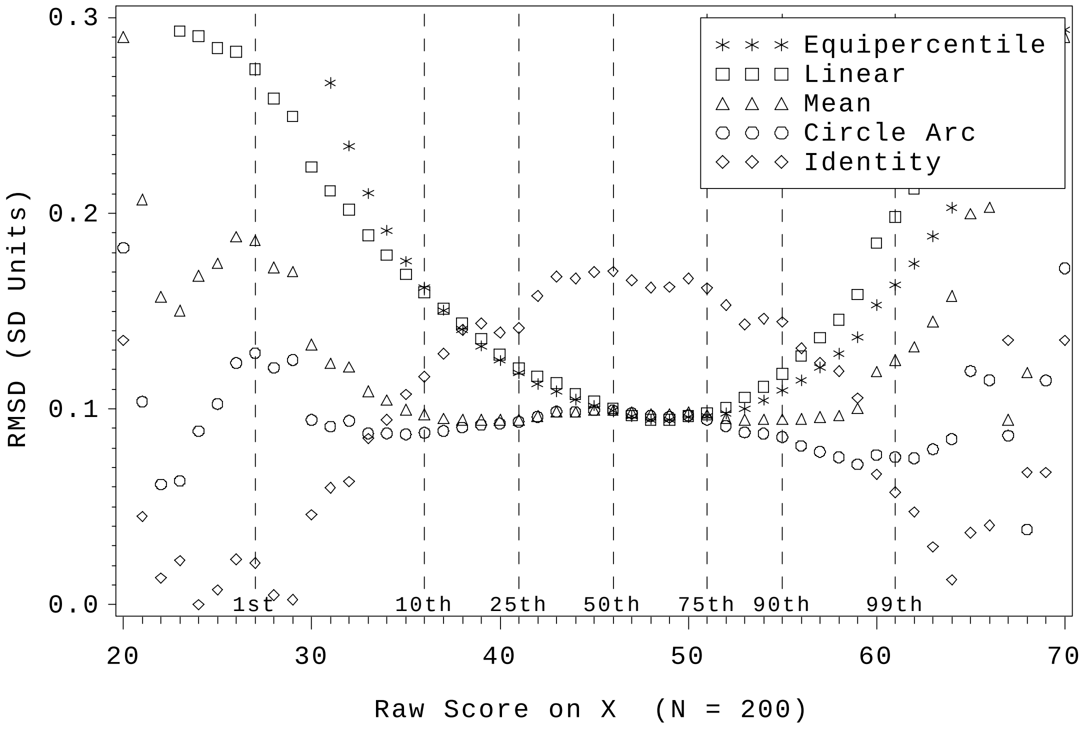

Figure A18. **Conditional root mean squared deviation: samples of 200, Test 4.**

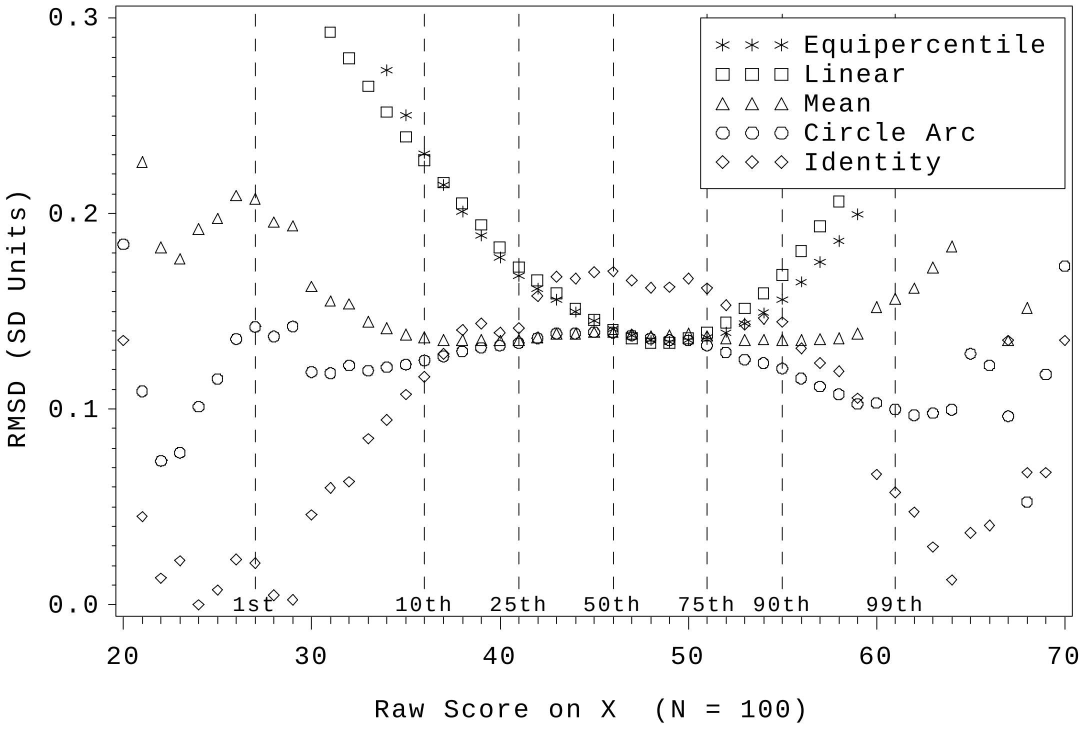

Figure A19. **Conditional root mean squared deviation: samples of 100, Test 4.**

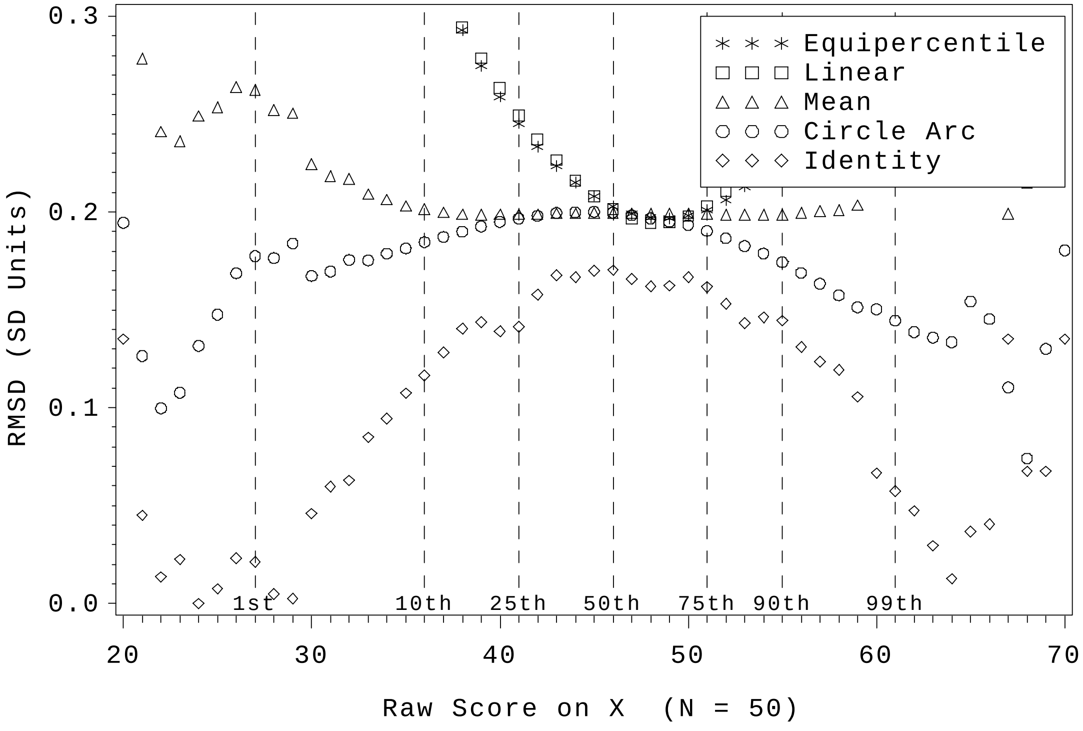

Figure A20. **Conditional root mean squared deviation: samples of 50, Test 4.**

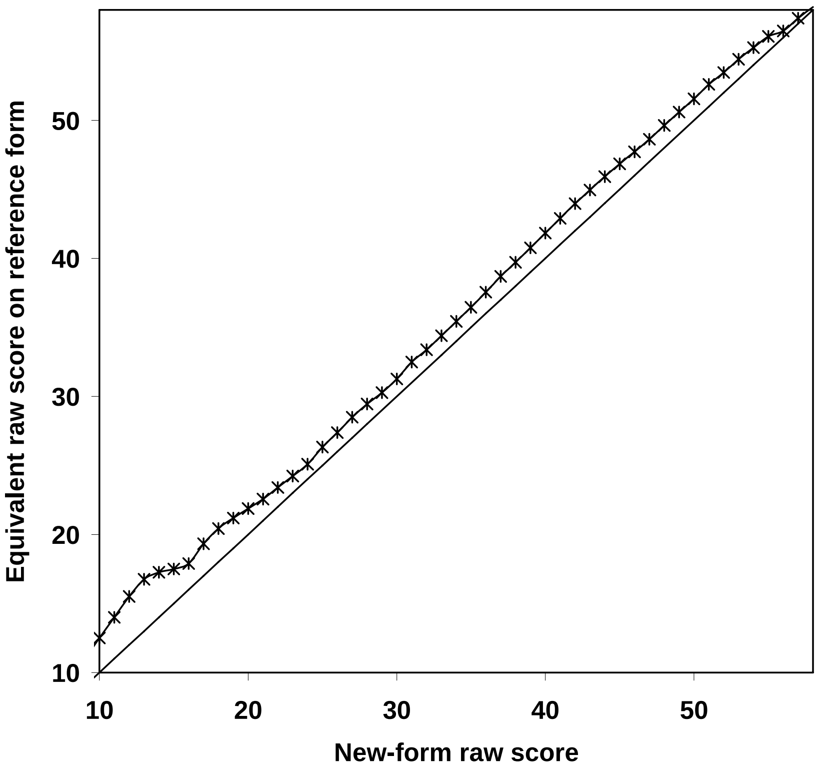

Figure A21. **Criterion equating for Test 5.**

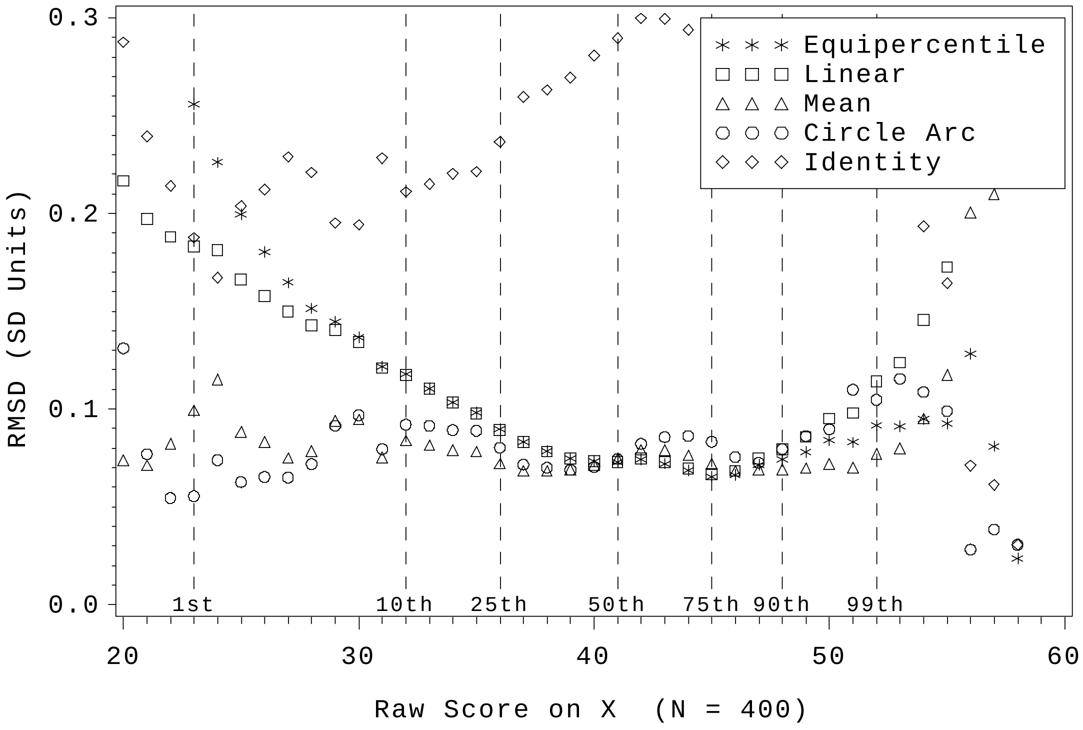

Figure A22. **Conditional root mean squared deviation: samples of 400, Test 5.**

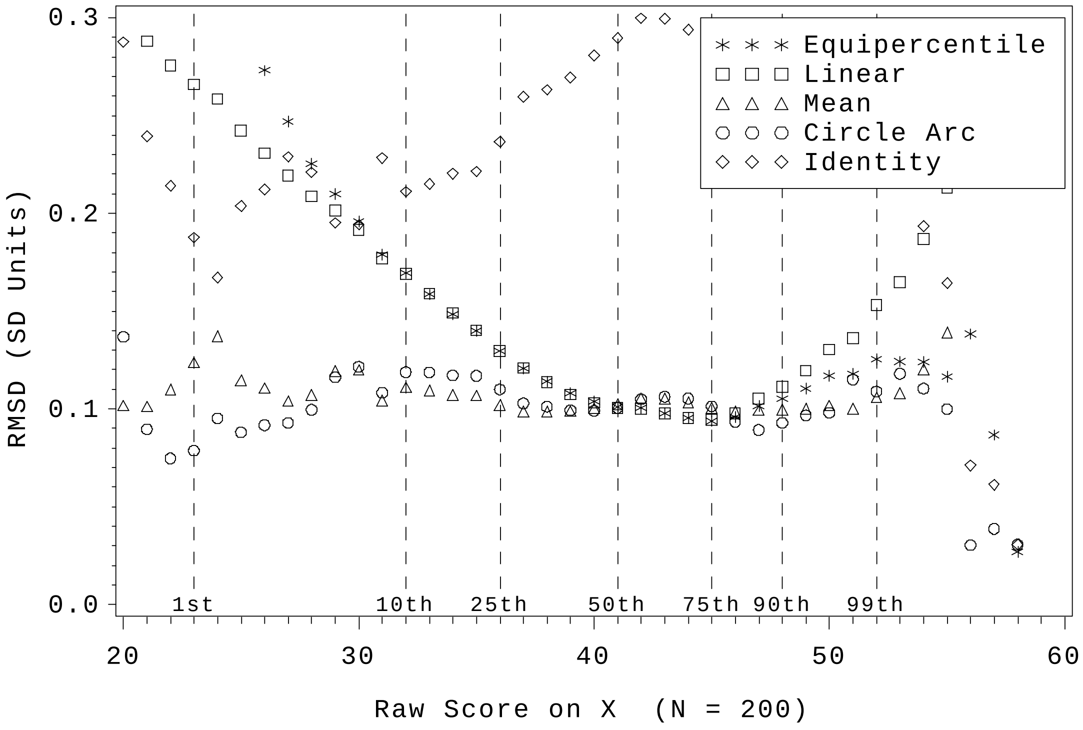

Figure A23. **Conditional root mean squared deviation: samples of 200, Test 5.**

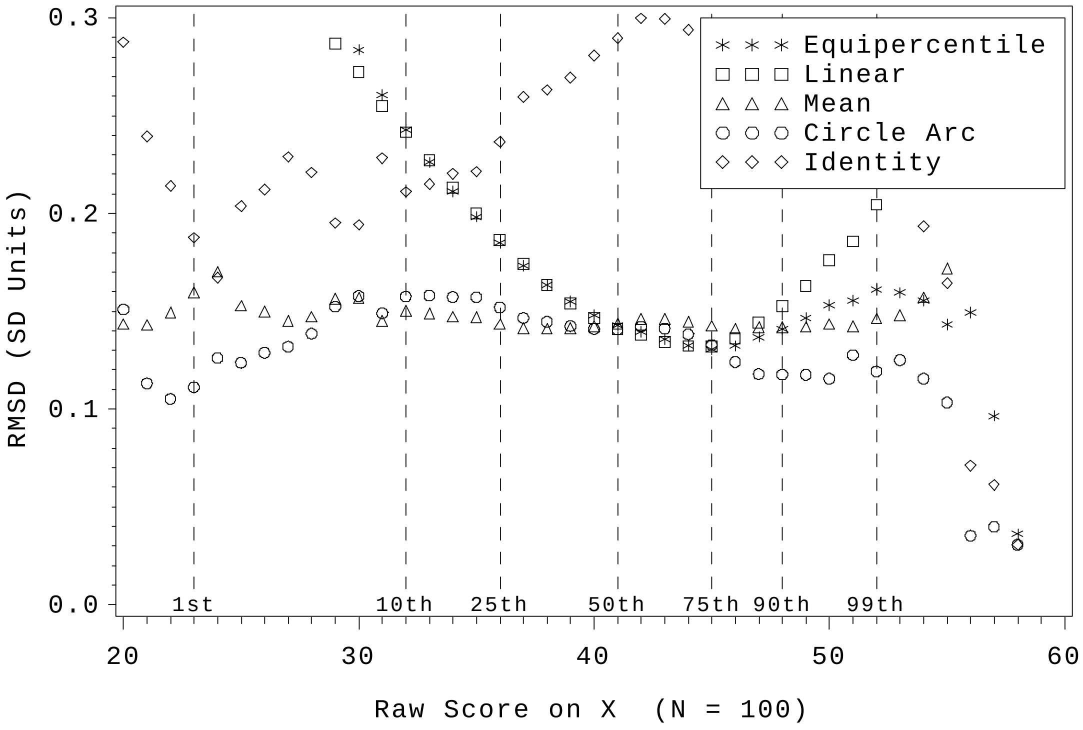

Figure A24. **Conditional root mean squared deviation: samples of 100, Test 5.**

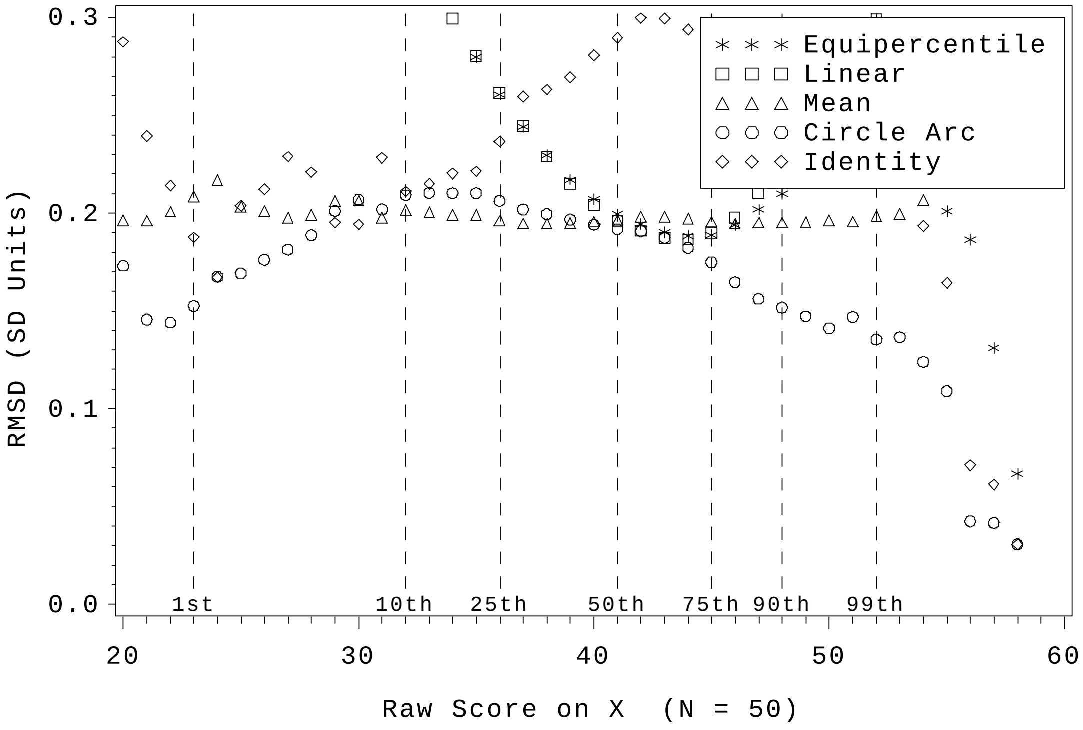

Figure A25. **Conditional root mean squared deviation: samples of 50, Test 5.**

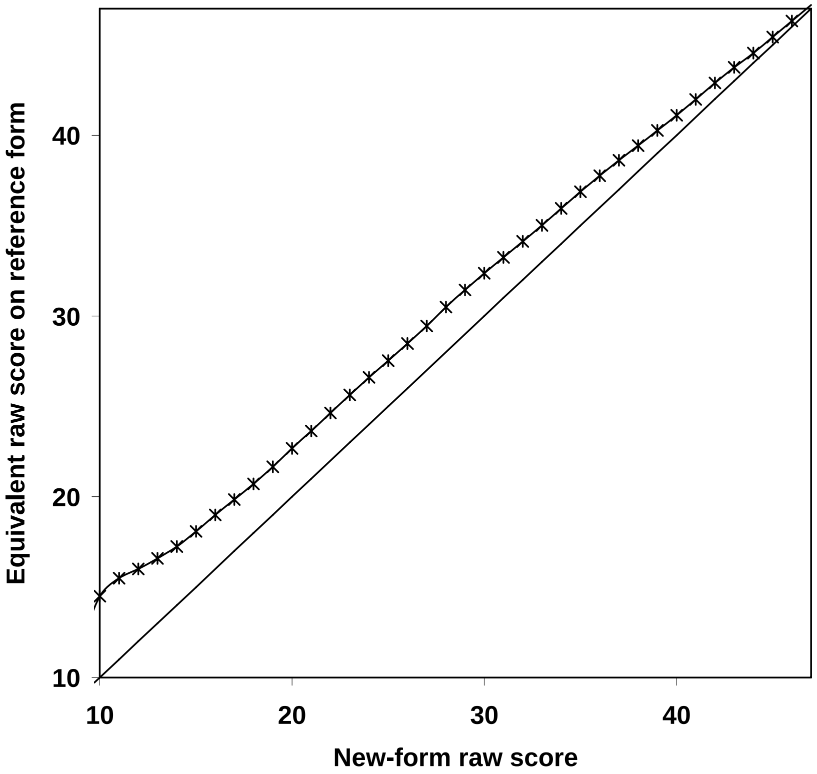

Figure A26. **Criterion equating for Test 6.**

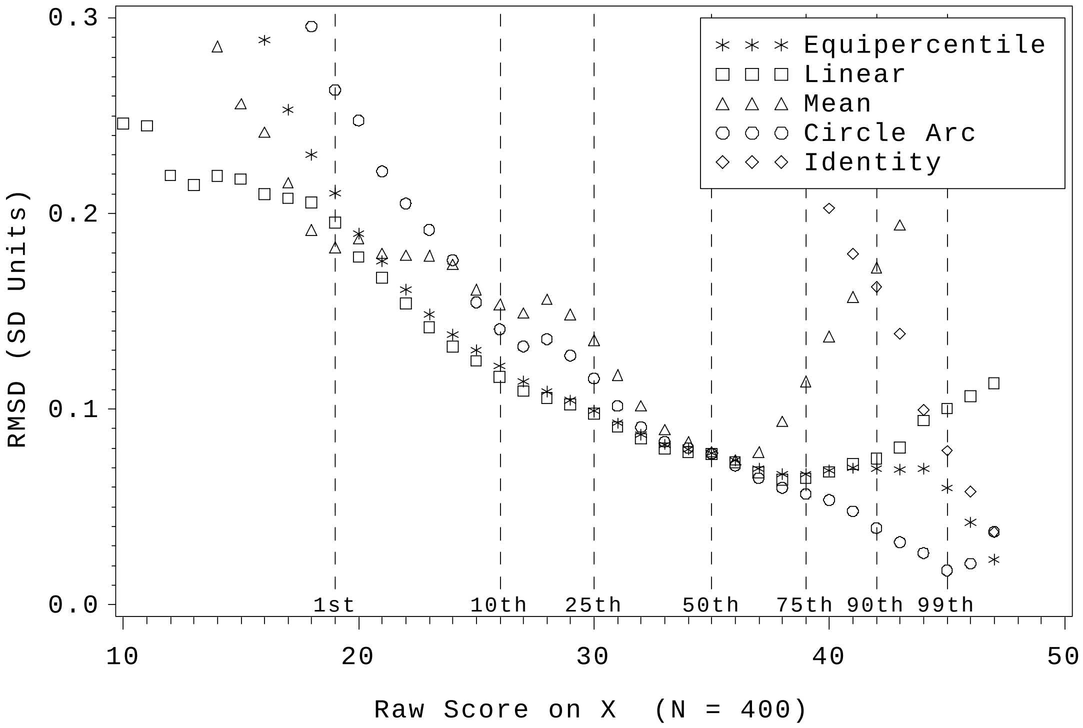

Figure A27. **Conditional root mean squared deviation: samples of 400, Test 6.**

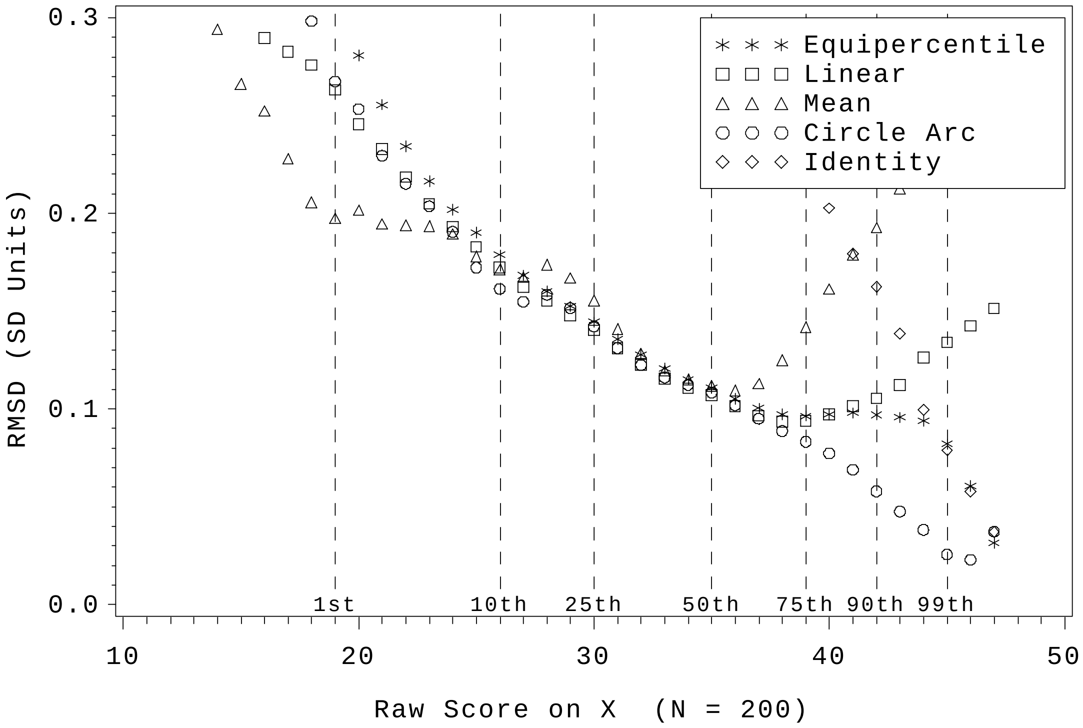

Figure A28. **Conditional root mean squared deviation: samples of 200, Test 6.**

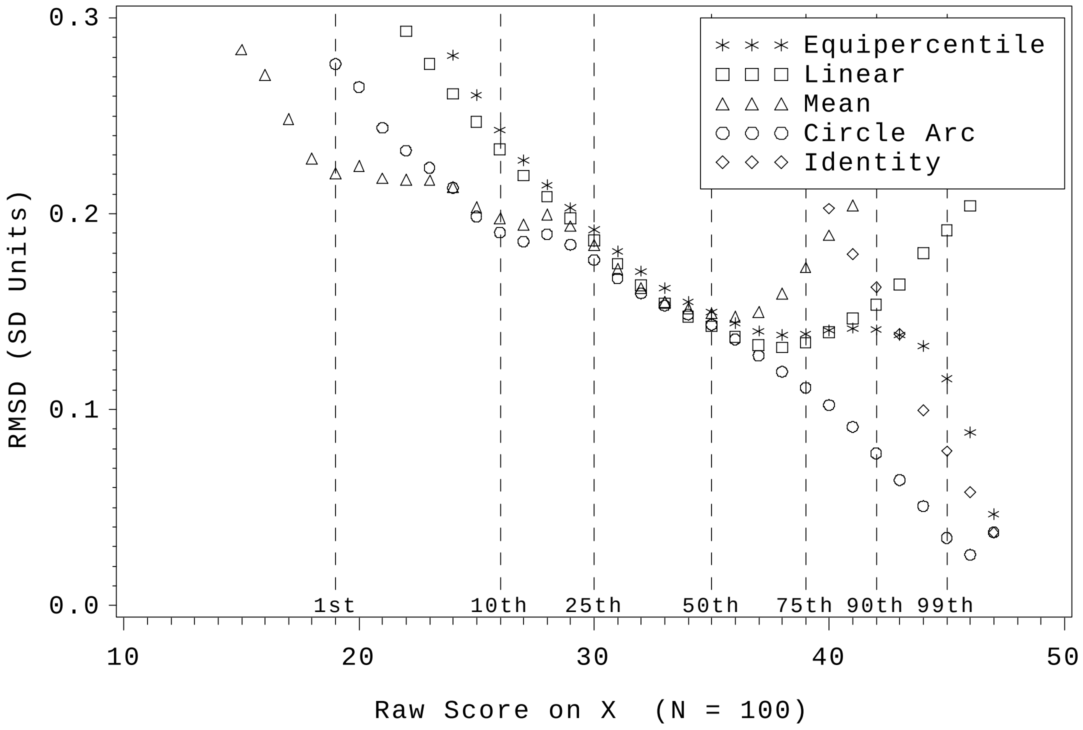

Figure A29. **Conditional root mean squared deviation: samples of 100, Test 6.**

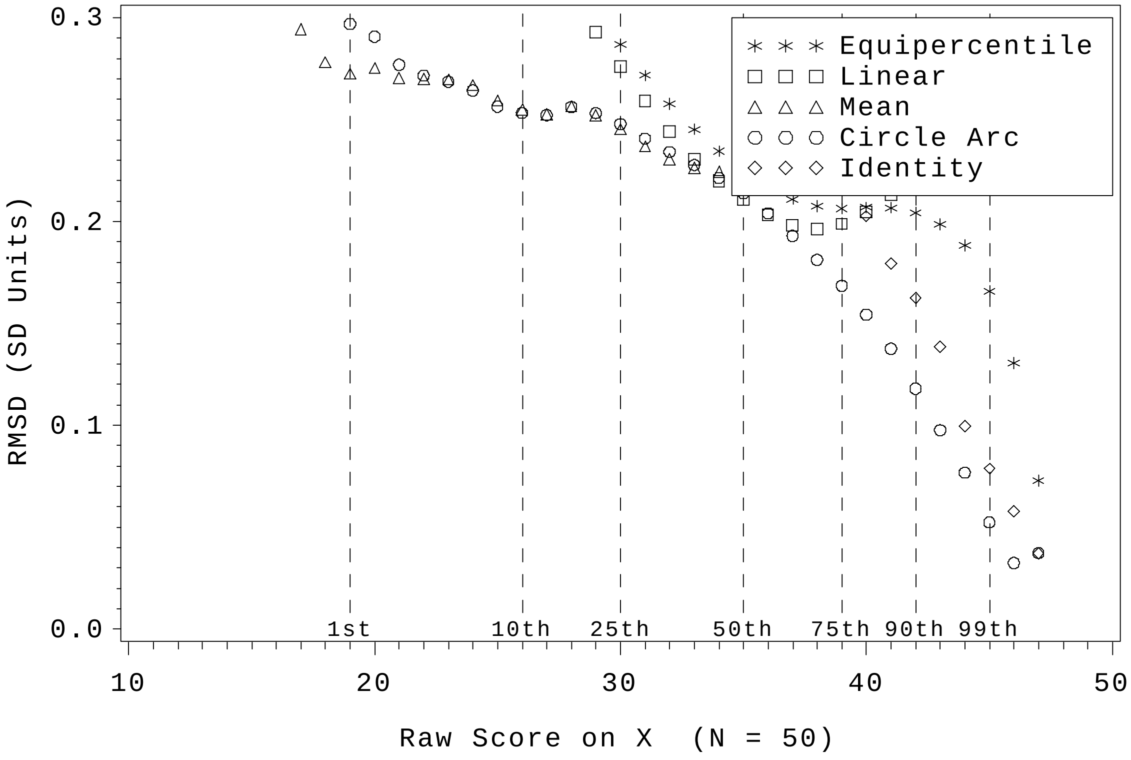

Figure A30. **Conditional root mean squared deviation: samples of 50, Test 6.**