Research Report

Investigating the Effectiveness of Collateral Information on Small-Sample Equating

Sooyeon Kim

Samuel A. Livingston

Charles Lewis

October 2008

ETS RR-08-52

Investigating the Effectiveness of Collatoral Information on Small-Sample Equating

Sooyeon Kim, Samuel A. Livingston, and Charles Lewis
ETS, Princeton, NJ

October 2008

Abstract

This paper describes an empirical evaluation of a Bayesian procedure for equating scores on test forms taken by small numbers of examinees, using collateral information from the equating of other test forms. In this procedure, a separate Bayesian estimate is derived for the equated score at each raw-score level, making it unnecessary to specify a parametric model for the equating function. Collateral information can come either from other forms of the same test or, possibly, from other tests having a similar structure. Our evaluation consisted of two resampling studies. Each study applied the Bayesian procedure to small samples drawn from large-sample data collected for an anchor equating. The large-sample equating function served as the criterion. The results of the two studies were somewhat inconsistent, leading to different conclusions regarding the use of the empirical Bayesian procedure with small samples.

Key words: Equating, empirical Bayes, collateral information, sample size, stability, bias

Acknowledgments

This study was funded by the 2007 Equating, Scaling, Applied Psychometrics Research Initiative. The authors would like to thank Neil Dorans, Gautam Puhan, Michael Walker, and Daniel Eignor for their helpful comments and Kim Fryer for the editorial help.

Table of Contents

Page

List of Figures

Page

Introduction

The Problem

In testing programs, multiple forms of a single test are used in different administrations, for test security purposes. Those forms must be as similar in difficulty as possible to ensure comparability of the scores across administrations. In practice, however, constructing test forms of equal difficulty is not easy, especially when the test items cannot be pretested. Statistical equating of the scores is necessary to overcome this problem. The purpose of equating is to establish an effective equivalence between scores on two test forms that are designed according to the same specifications. 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.

Some testing programs, including many certification tests, are administered in low volume situations. Thus, it is often hard to obtain data from as many as 50 examinees for test equating. Nevertheless, these programs need to provide, in a timely manner, comparable scores across different administrations and test forms. Accordingly, research on equating with small amounts of data is necessary to establish practical guidelines for this endeavor. The present study was designed to explore the effectiveness of empirical Bayes (EB) estimation, incorporating collateral information to improve the accuracy of equating with small data sets.

Previous Attempts

Estimated equating relationships include estimation error. Random equating error, which is typically indicated by the standard error of equating (SEE), is present whenever samples are used to estimate equating relationships in populations (Kolen & Brennan, 2004). Sample size has a direct effect on the SEE, which is the measure of the sampling variability of estimated equating functions. Some experts believe that use of the identity function (i.e., assuming the difficulty of the forms to be exactly equal) is often preferable to equating with extremely small samples of test takers, as the large random equating error associated with very small samples negates the benefits of equating (Harris, 1993; Kolen & Brennan). While this solution completely eliminates equating error due to sampling, it introduces a bias if the test forms are not of equal difficulty. There are a few empirical studies on the impact of equating with small samples with respect to equating error or bias (Kolen & Whitney, 1982; Parshall, Du Bose Houghton, & Kromrey, 1995; Skaggs, 2005).

Parshall, Du Bose Houghton, and Kromrey (1995) examined the effects of sample size on the stability and bias of linear equating of two parallel forms, in a nonequivalent groups with anchor test (NEAT) equating design, with samples ranging from 15 to 100. Their results suggested negligible levels of equating bias even with small samples. However, SEEs increased substantially as sample size decreased. Sampling error was smallest in proximity to the mean raw score of the new form to be equated. The SEE increased monotonically, but not linearly, as a function of the scores' deviation from the mean.

Skaggs (2005) studied the equating of the passing score on a certification test using linear and nonlinear methods in an equivalent-groups design, with samples ranging from 25 to 200 examinees. The SEE became smaller, but equating bias changed little, as sample size increased. Even when the sample included 200 examinees, substantial equating error occurred on at least part of the raw score scale. As a result, a significant percentage of examinees were misclassified by their pass/fail designations. Skaggs found that equating with samples as small as 25 observations may produce a large amount of equating error. Equating under these circumstances may be worse than not equating at all.

Studies have been conducted to explore potential methodological solutions to equating with small samples (Kim, von Davier, & Haberman, 2006, 2007; Livingston 1993). Livingston (1993) examined the effectiveness of presmoothing the score distributions on the accuracy of chained equipercentile equating with samples of 25, 50, 100, and 200 examinees per form. Using log-linear models to presmooth the distributions, he found that presmoothing improved equating accuracy about as much as doubling the sample size. Log-linear presmoothing might yield a smaller SEE, but it also might bias the equating results.

Recently, Kim, von Davier, and Haberman (2006, 2007) proposed the synthetic function method for equating with small samples. The synthetic function is essentially a weighted combination of a sample equating function and the identity function, using a prespecified weight system. Because the SEE for the identity function is zero, the SEE for the synthetic function can be substantially reduced, as compared with the SEE of the sample equating function. The use of the identity function was derived from the idea that any increase in systematic error that the identity function produces is more than offset by the decrease in random error. Using a variety of real data sets, Kim and associates examined linking bias and error among the identity, mean, chained linear, and synthetic functions, with samples ranging from 10 to 200. In general, the synthetic function performed better than did traditional linear functions for samples of fewer than 30 examinees. The benefits of the synthetic function, however, depended heavily on the extent to which forms differed in difficulty. When test forms were nearly parallel, the synthetic function exhibited less linking error and only a small degree of bias. Conversely, when the forms clearly were not parallel, the effectiveness of the synthetic function was diminished due to the extensive bias that the identity function caused. In addition, a major problem with the synthetic function is that the choice of weights for the identity and the sample equating functions is unclear.

Using Collateral Information

Some researchers have investigated the use of collateral information in IRT calibration (Mislevy, Sheehan, & Wingersky, 1993) and in DIF analysis (Sinharay, Dorans, Grant, Blew, & Knorr, 2006), but we found only one reference describing its use in estimating a score equating relationship under the IRT equating framework (Hsu, Wu, Yu, & Lee, 2002). Recently, Livingston and Lewis (2007) proposed an EB estimation procedure, which may have the potential to improve the accuracy of equating with small amounts of data. Although the use of collateral information to improve the accuracy of estimates is not a new approach, no studies have been conducted to investigate its use with samples of fewer than 50 examinees per form under the classical equating framework. The goal of the present study is to explore the effectiveness of using collateral information in such situations. Prior equatings can be selected from either the same test or from other tests having similar content structures or specifications. In practice, the latter procedure may be more useful than the former because, for many low-volume tests, multiple forms may not be available.

An equating function computed by the EB approach is essentially a compromise between a sample equating function and prior equatings. In our EB procedure, this combination is computed separately for each possible raw score on the targeted new form (i.e., the new form to be equated).

In mathematical notation, the equated score, the reference-form raw score corresponding to a given raw score on the new form, is given by

$$\hat{y}_{EB} = \frac{\frac{1}{\sigma^2_{prior}} y_{prior} + \frac{1}{\sigma^2_{eq}} y_{eq}}{\frac{1}{\sigma^2_{prior}} + \frac{1}{\sigma^2_{eq}}}, \tag{1}$$

$$\sigma^2_{prior} = \frac{1}{m-1}\sum_{i=1}^{m}\left(y_i - y_{prior}\right)^2 - \frac{1}{m}\sum_{i=1}^{m} CSEE^2_{y_i}, \tag{2}$$

where *i* indexes the prior equatings, 1 through *m*, and y_i is the new-form raw score value for which the equated score is to be found. As shown in Equation 1, the weight for the equated score, as determined by the current equating (y_{eq}), is the inverse of its variance, i.e., of the squared conditional SEE (σ^2_{eq}). Therefore, y_{eq} will receive a greater weight as the size of equating samples (particularly the smaller sample, which is usually the new form sample) in the current equating increases. The weight for the equated score implied by the prior equatings (y_{prior}) is a function of its variability across the prior equatings and the degree of stability of each prior equating ($CSEE^2_{y_i}$), as shown in Equation 2. This function gives y_{prior} a greater weight when the prior equatings are less variable. The current equating is included in the set of equatings used to estimate the prior mean and variance, because the prior mean and variance are estimates for a domain of equatings, and the current equating is a member of that domain.[1] The inclusion of the current equating in the estimation of the prior mean and variance has the effect of increasing its influence on the posterior equated score ($\hat{y}_{EB}$), particularly when only a few prior equatings are available.

This point-by-point EB procedure does not require the equating transformation to follow a particular mathematical form. It treats the determination of the equated score at each point as a separate estimation problem to be solved by incorporating information from prior equatings. Bayesian estimates are derived separately for each equated score, rather than for the parameters of the equating function, allowing the procedure to estimate nonlinear equating functions without assuming that they have a particular mathematical form. Extensions of the procedure permit the use of collateral information from the equating of test forms having different numbers of items,[2] by expressing the raw score on each form as a percentage of the maximum possible raw score on that form (i.e., percent-correct). If the new form in a prior equating has a different number of items than the targeted new form, the percent-correct scores possible on the targeted new form will be different from those possible on the prior form. In that case, interpolation is necessary to determine the equated scores in the prior equating. The general EB procedures are presented in the appendix.

Because this EB procedure does not assume a particular mathematical form for the equating transformation, it can be used with any equating method, or any combination of equating methods. It allows for the current equating to be done by one method, some of the prior equatings by another, and other prior equatings to be done by yet another. The present study took advantage of this flexibility. Two different linear methods were used for the current equating.[3] One of these methods was chained linear equating, in which the equating function is the composition of the equating of the new-form score to the anchor score in the new-form examinee sample and the equating of the anchor score to the reference-form score in the reference-form examinee sample. The other method was mean equating; equating by simply adding the same quantity at all points of the new-form raw-score scale. In this study, the quantity to be added was determined by the chained linear method.[4] The estimates of the conditional error of standard equating (CSEE) of the current equating were computed by the delta method, for both the chained linear equating (see Kolen & Brennan, 2004, pp. 254-255) and the chained mean equating (see Braun & Holland, 1982, p. 36). Some of the prior equatings had been done by the chained linear method; while others had been done by applying the chained equipercentile method to pre-smoothed score distributions.

The Evaluation

The purpose of the present study was to determine empirically how much the accuracy of small-sample equating could be improved by using EB methods that incorporate collateral information from prior equatings. To determine the accuracy of a small-sample equating, it is necessary to know (or to estimate accurately) the results of equating in the population that the small sample represents. The empirical evaluation of the Bayesian equating procedure consisted of two resampling studies. In each study, the Bayesian equating procedure was applied to small samples drawn from large groups of examinees who had taken the tests in a

NEAT equating design. The large-group equating results served as a target equating for the small-sample equatings.

The sample sizes for the new form in the small-sample equatings ranged from 10 to 200 examinees. We equated the new form to the reference form in each pair of small samples, by each equating method (chained linear equating and mean equating), with and without collateral information from previous equatings. One additional equating procedure, robust but not truly Bayesian, was also applied to the data from each replication. This procedure was a synthetic function formed by averaging the results of the chained linear equating with the identity transformation, weighting these two functions equally. The results of the small-sample equatings by these five methods were compared with the results of the original large-group equating, which served as the criterion equating.

The evaluation of the accuracy of the small-sample equatings was based on the root-mean-squared error (RMSE) over repeated sampling, where *error* is defined as the difference between the equated scores produced by the small-sample equating and the large-group criterion equating. The RMSE can be decomposed into two orthogonal components: (a) the deviation of the individual small-sample results from their average value (i.e., the SEE), and (b) the deviation of this average small-sample value from the large-group value (i.e., the statistical bias of the procedure). These statistics were computed at each new-form raw-score value, because an equating procedure can be extremely accurate at some score values and extremely inaccurate at other score values. The formulas are shown in Equations 3 to 5:

$$Bias_i = \overline{d}_i = \frac{\sum_{j=1}^{J}[\hat{e}_j(x_i) - e(x_i)]}{J}, \tag{3}$$

$$SEE_i = s(d_i) = \sqrt{Var_j\left[\hat{e}_j(x_i) - e(x_i)\right]} = \sqrt{Var_j\left[\hat{e}_j(x_i)\right]}, \tag{4}$$

$$RMSE_i = \sqrt{\overline{d}_i^2 + s^2(d_i)}, \tag{5}$$

where *i* indexes the raw scores on the targeted new form, *j* indexes the replications of the procedure, *J* is the total number of replications (500), $\hat{e}_i(x_i)$ is the equated score at raw score *x*, and $e_i(x_i)$ is the criterion equating function. To compare the accuracy of the equating methods in the full examinee population, these statistics were averaged over the new-form raw-score distribution, weighting the "error" at each new-form raw-score value in proportion to the frequency of that raw score in the group of examinees taking the new form, in the data set for the large-group equating. Using w_i to represent the proportional frequency at score *i*, the resulting statistics were: (a) the weighted root mean squared bias,[5] $\sqrt{\sum_i w_i Bias_i^2}$, (b) the weighted SEE, $\sqrt{\sum_i w_i SEE_i^2}$, and (c) the weighted RMSE, $\sqrt{\sum_i w_i RMSE_i^2}$.

In Studies 1 and 2, the large data sets included more than 6,000 examinees per form, and the prior equatings were other forms of the same test as the targeted new form. Two factors that were systematically varied in these studies were; (a) the number of prior equatings used as collateral information, and (b) the number of examinees in the new form sample. The number of prior equatings ranged from 3 to 12, and the sample sizes for the new form ranged from 10 to 200 examinees.

Study 1

Data

The data for Study 1 came from two national administrations of a licensure test for teachers. The test assesses the basic understanding of curriculum planning, instructional design, and assessment of student learning in the elementary grades. The test form administered in April 2004 will be referred to as Form *X*, and the test form to which it was equated, administered in March 2004, will be referred to as Form *Y*. The 6,019 examinees tested in April 2004 will be referred to as Population *P*; the 6,386 examinees tested in March 2004 will be referred to as Population *Q*. Forms *X* and *Y* each consisted of 108 multiple-choice (MC) items,[6] of which 36 were anchor items for equating. The anchor score will be indicated by *V*. Descriptive statistics for these two national administrations are shown in Table 1. The mean anchor scores of the two populations differed by 0.19 items correct, an effect size of only 0.05. The internal consistency reliabilities for both tests (.84 and .85) and anchors (.67) were moderate. The high correlation between the test scores and anchor scores (.88) reflects the inclusion of the anchor items in both scores.

Table 1

Descriptive Statistics for the Observed Distributions of X, V in P and Y, V in Q: Study 1

	N	μ	σ	SEM	Reliability	ρ
X		78.58	10.78	4.2	.84	
V_P	6,019	26.75	4.08	2.3	.67	.88
Y		79.44	10.85	4.2	.85	
V_Q	6,386	26.56	4.06	2.3	.67	.88

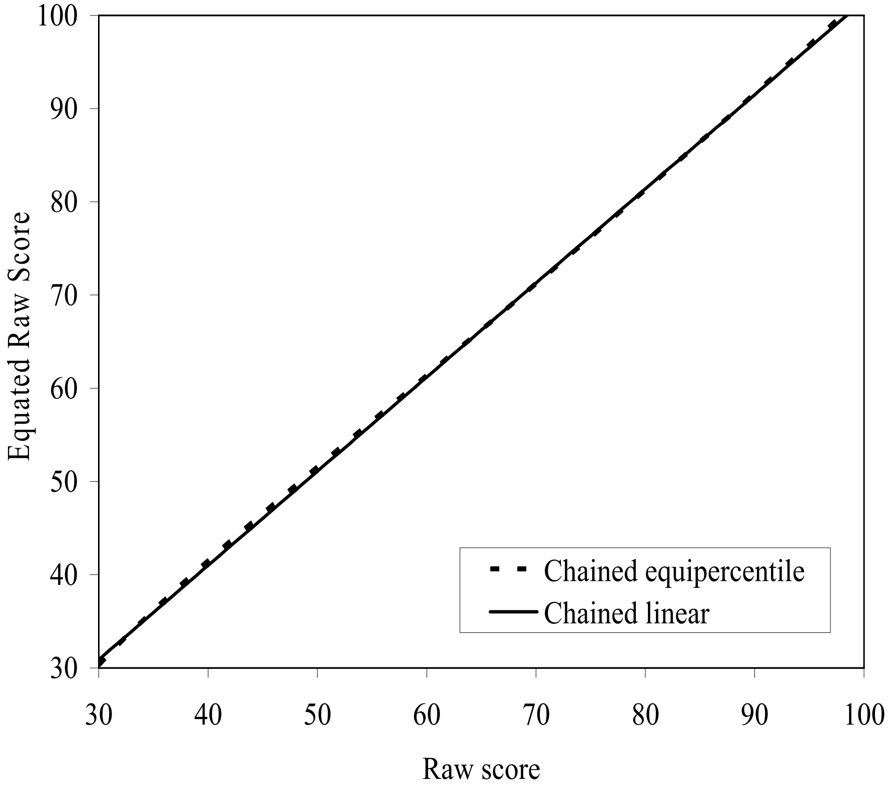
Note. SEM = standard error of measurement, ρ = correlation between total score and anchor.

The criterion equating function was determined by an equating that made use of all the available data. This equating was computed by both the chained linear and chained equipercentile methods. Figure 1 plots the raw-to-equated-raw score conversion lines produced by the chained linear and chained equipercentile methods, in the portion of the score scale where any examinees' scores were observed. The difference was consistently less than half a raw-score point, the quantity that Dorans and Feigenbaum (1994) described as the *difference that matters*. Based on those observations, the chained linear equating function was selected as the criterion for the small-sample equatings in this study.

Procedure

Resampling. The study included 500 replications of the small-sample sampling/equating procedure at each of six specified sample sizes for the new-form equating sample: 10, 25, 50, 75, 100, and 200. The reference-form equating sample size was 200 in every case, but samples differed in every replication. In real test equating situations, there is often the possibility of enlarging the reference-form equating sample by combining data from more than one administration of that form. There is no such possibility for the new form. The examinees for 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 Population *P* and a reference-form sample from Population *Q*.
2. Equate Form *X* to Form *Y* in those samples, by the chained linear method and by chained mean equating.



Raw score	Frequency
0-30	1
31-36	6
37-42	14
43-48	38
49-54	107
55-60	210
61-66	410
67-72	801
73-78	1158
79-84	1328
85-90	1196
91-96	627
97-102	122
103-108	1
Total	6019

***Figure 1*. Plot of raw-to-equated-raw score for chained linear and chained equipercentile, and frequency distribution of new form *X* scores in total group for Study 1.**

3. Use those two equatings as the current equatings in the EB estimation procedure, to determine the two corresponding EB estimations.
4. Compute the synthetic linking function, by averaging the chained linear equating and the identity function, weighting them equally.

Collateral information. The collateral information for Study 1 consisted of the equatings of 12 previous forms of the test, using either the chained linear method or the chained equipercentile method, with sample sizes that ranged from 300 to 6,000 examinees. The number of prior equatings used as collateral information was systematically varied, with levels of 3, 6, 9, and 12 prior equatings.[7] Where fewer than 12 prior equatings were used, the equatings to be used were selected at random from the 12 available equatings without replacement. The random selection of prior equatings was made separately and independently for each replication of the EB procedure.

Results

Figures 2 to 7 show the agreement between the results of each small-sample equating method (from chained linear, chained mean, EB chained linear, EB chained mean, and synthetic) and those of the large-group equating (chained linear), computed separately at each new-form raw-score level within the range of scores observed in the large group. The strength of the agreement is indicated by the RMSE over the 500 replications; a value of zero would indicate perfect agreement. (Figures A1 to A6 in the appendix show similar plots of the conditional bias and the conditional SEE.) Each figure includes a separate curve for each of the five small-sample methods. For the two EB methods, only the results incorporating all 12 prior equatings are shown. The dashed vertical lines indicate z-score values of -2, -1, 0, +1, and +2, in the large group.

Figure 2 shows the results for the new-form samples of only 10 examinees. With this very small new form sample size, the methods with the smallest RMSE values throughout most of the score range were the EB chained linear method and the synthetic function (the equally-weighted average of the chained linear equating and the identity). Because the test forms being equated did not differ greatly in difficulty, the synthetic function performed well; slightly better than the EB chained linear method for raw scores within about 1.5 standard deviations (SDs) of the mean, but not as well for higher and lower scores. Both EB methods clearly outperformed their non-EB counterparts. The EB chained linear method outperformed EB chained mean equating, with a much smaller RMSE in the denser parts of the score distribution and only a slightly larger RMSE for scores more than 1.5 standard deviations from the mean.

Figure 3 shows the conditional RMSEs for the new-form samples of 25 examinees. The comparisons between methods are similar to those for samples of 10 examinees, but the differences between the EB methods and their non-EB counterparts are much smaller. Figure 4 shows the conditional RMSEs for the new-form samples of 50 examinees. The most striking results were the relatively poor performances of the two mean equating methods in the denser portion of the score distribution and of the chained linear method for scores more than one SD below the mean. The EB methods performed about as well as their non-EB counterparts in the middle of the distribution and better, particularly for the chained linear method, for scores more than 1 SD from the mean. The synthetic function performed slightly better than the EB chained linear method.

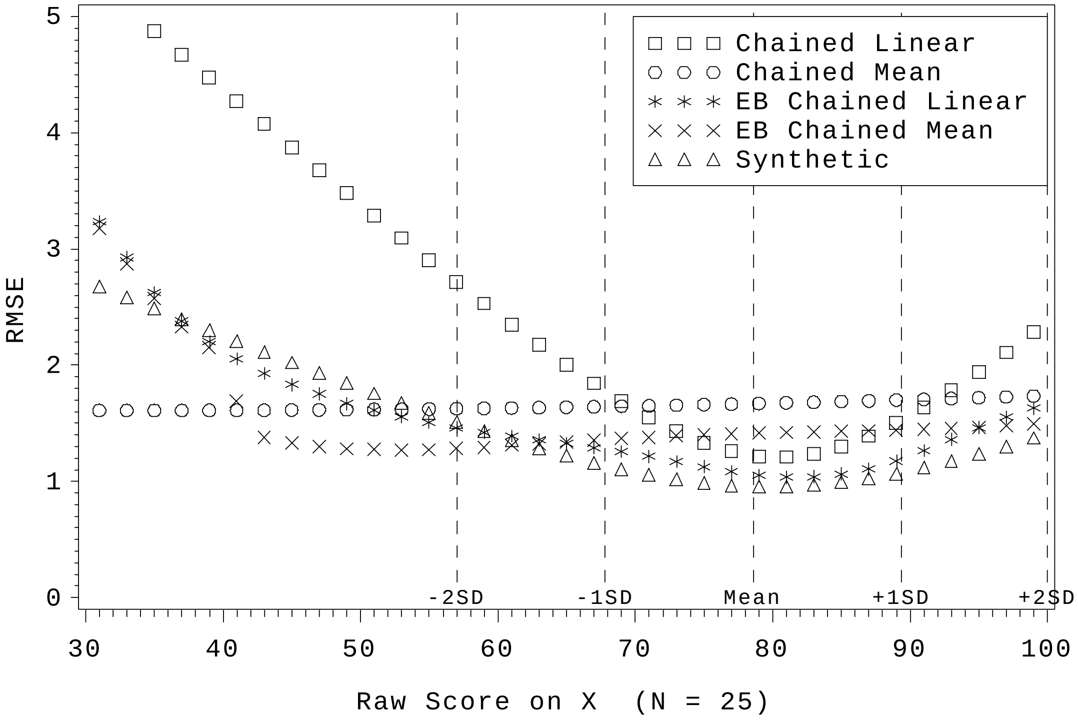


***Figure 2.* Conditional root mean squared error at samples of 10 in Study 1.**

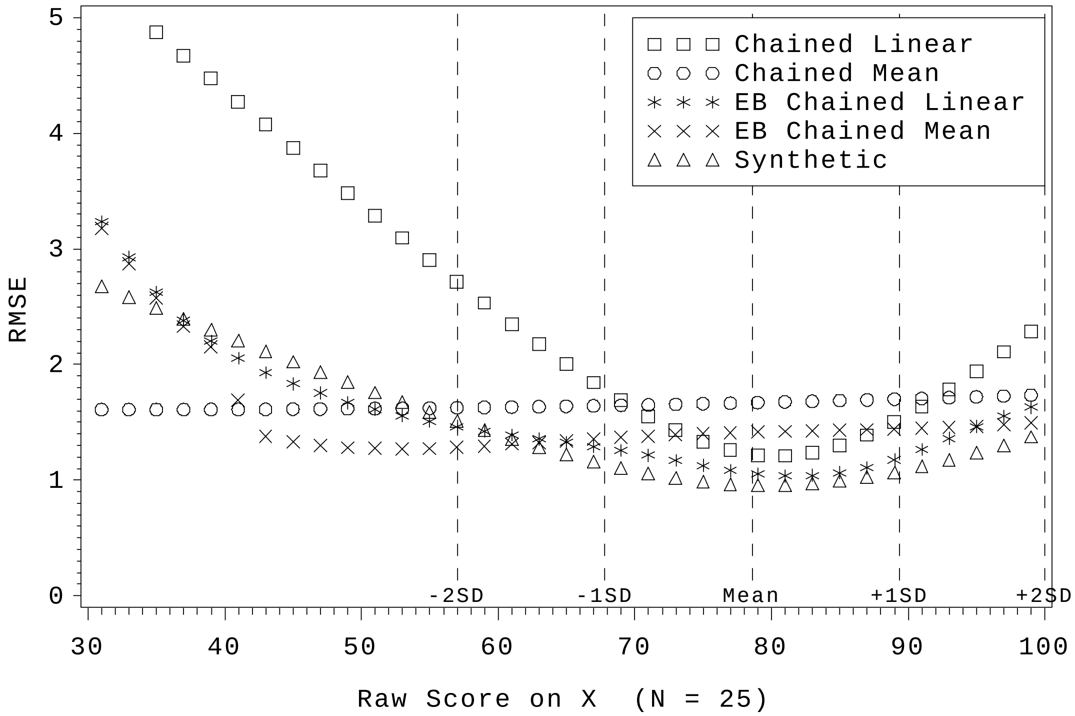


Figure 3. **Conditional root mean squared error at samples of 25 in Study 1.**

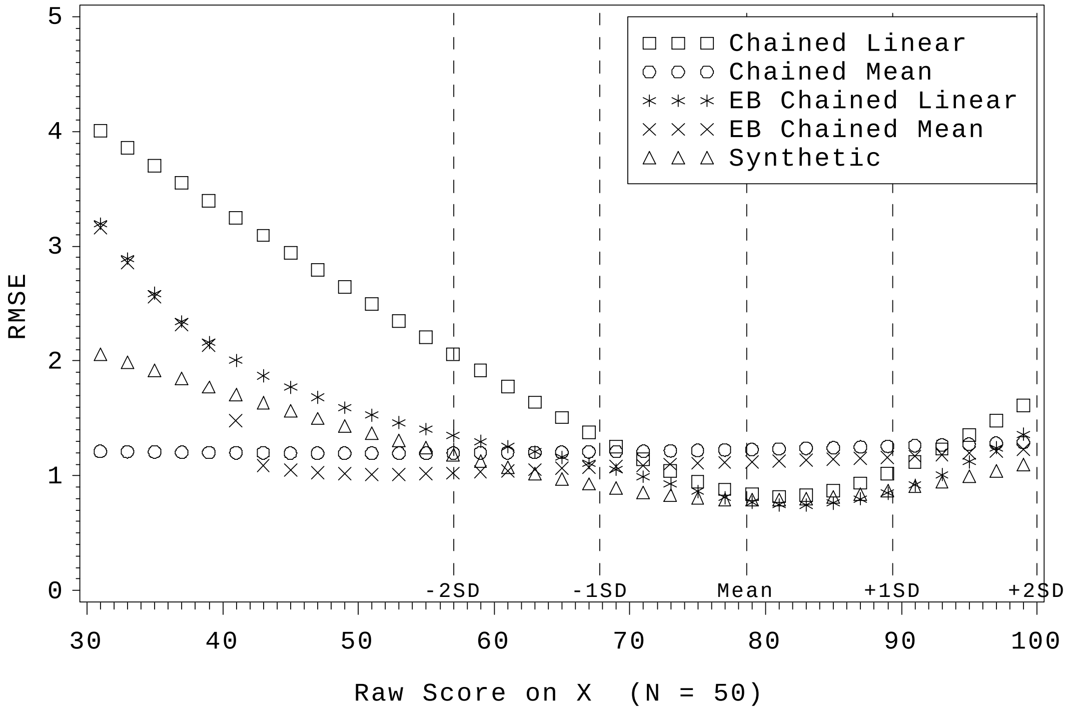


Figure 4. **Conditional root mean squared error at samples of 50 in Study 1.**

Figures 5 and 6 show plots of the conditional RMSEs for the new-form samples of 75 and 100 examinees, respectively. The comparisons between methods are similar to those for samples of 50 examinees; except for the synthetic function, which showed a slightly larger RMSE than both the EB and non-EB chained linear equatings in the middle of the distribution. Figure 7 shows the conditional RMSEs for the new-form samples of 200 examinees. With samples of this size, the EB methods differed only slightly from their non-EB counterparts, because the prior information had only a small influence on the estimated equating relationship. The two chained linear methods had much lower RMSEs than the other methods, in the middle of the distribution.

Figures 8 to 13 show the RMSE separated into its two orthogonal components (bias and standard error) and averaged over score levels, weighting each score level by its frequency in Population *P*. Each figure contains a single data point for each of the non-EB methods and four data points for each EB method, corresponding to the different numbers of prior equatings used as collateral information (3, 6, 9, or 12). The horizontal position of the data point indicates the bias in the results, indicated by the mean (over the 500 replications) of the differences between the small-sample equated score and the large-group criterion equating, averaged over score levels. In averaging over score levels, each score level is weighted by its population frequency.

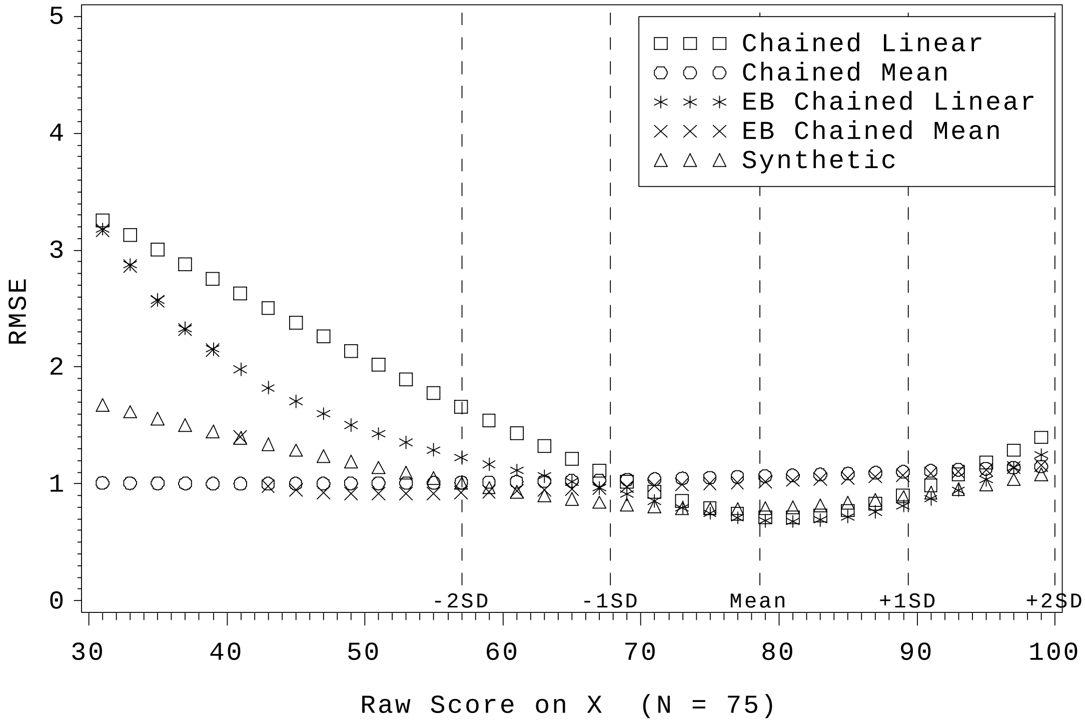


Figure 5. **Conditional root mean squared error at samples of 75 in Study 1.**

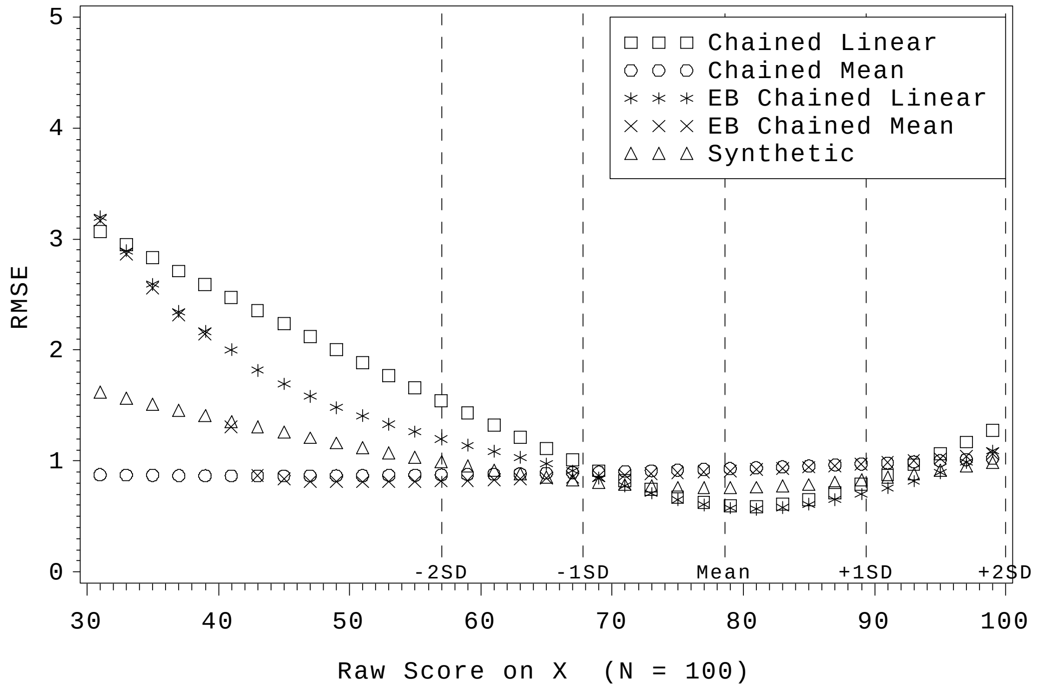


Figure 6. **Conditional root mean squared error at samples of 100 in Study 1.**

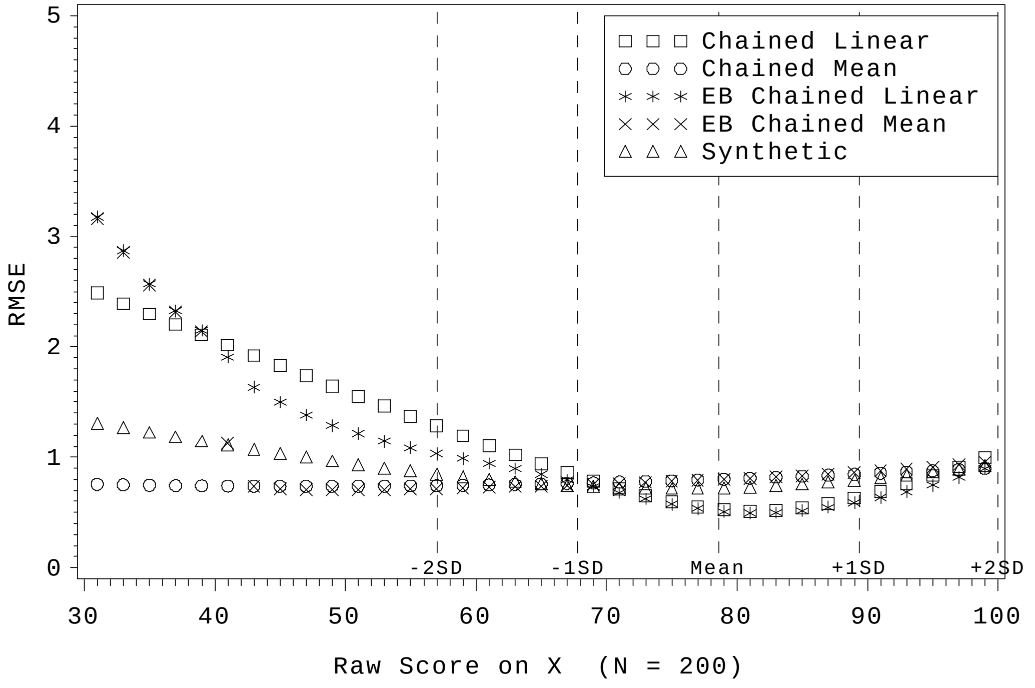


Figure 7. **Conditional root mean squared error at samples of 200 in Study 1.**

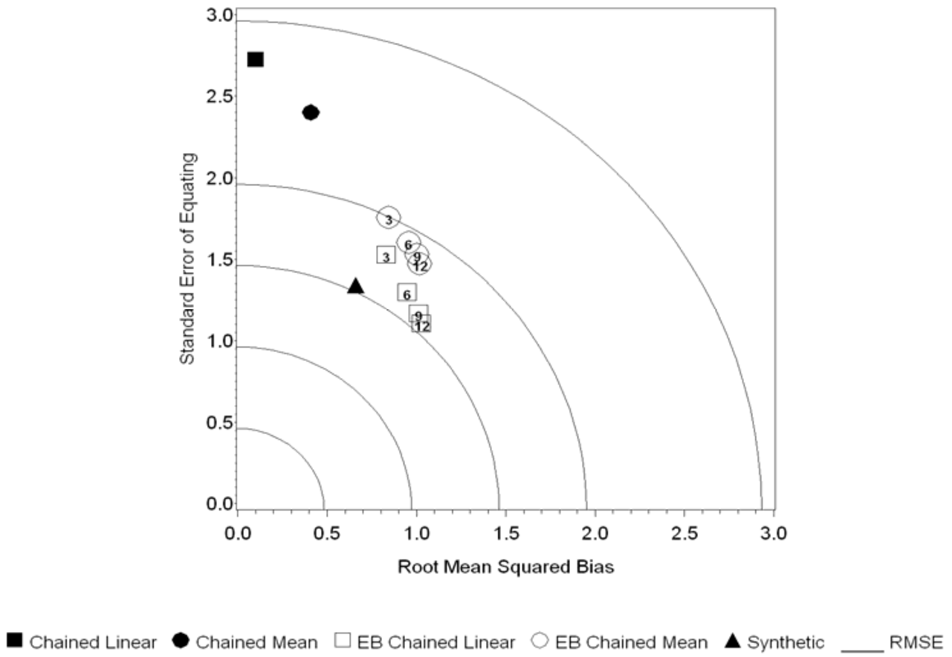


Figure 8. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 10 in Study 1.**

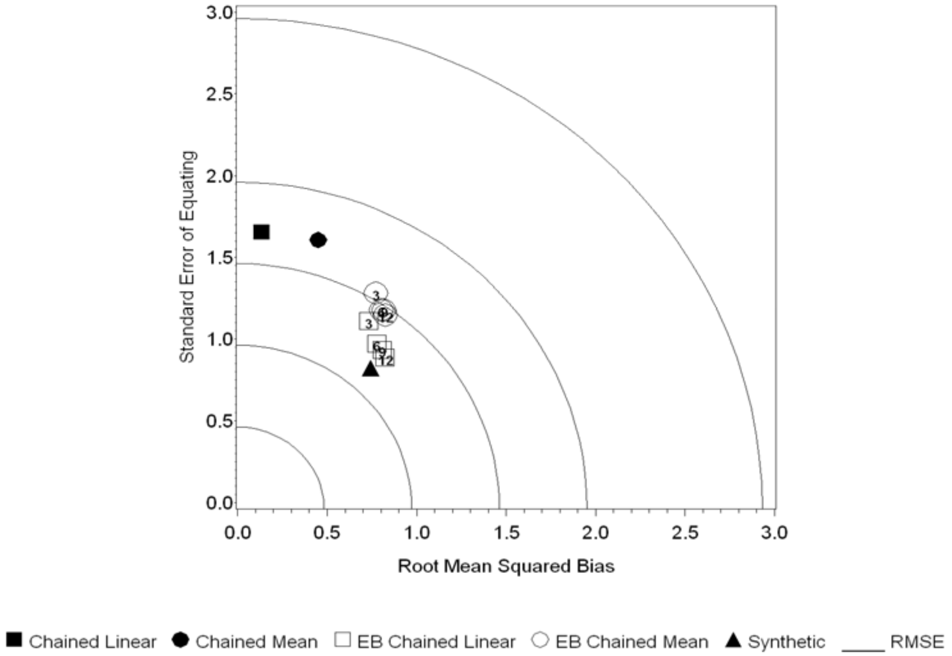


Figure 9. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 25 in Study 1.**

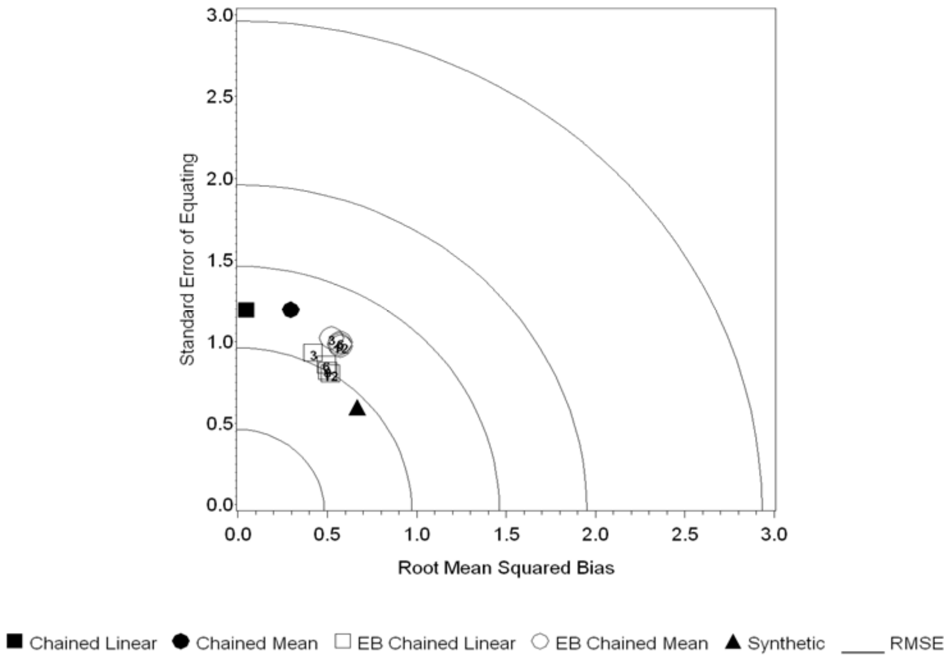


***Figure 10.* Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 50 in Study 1.**

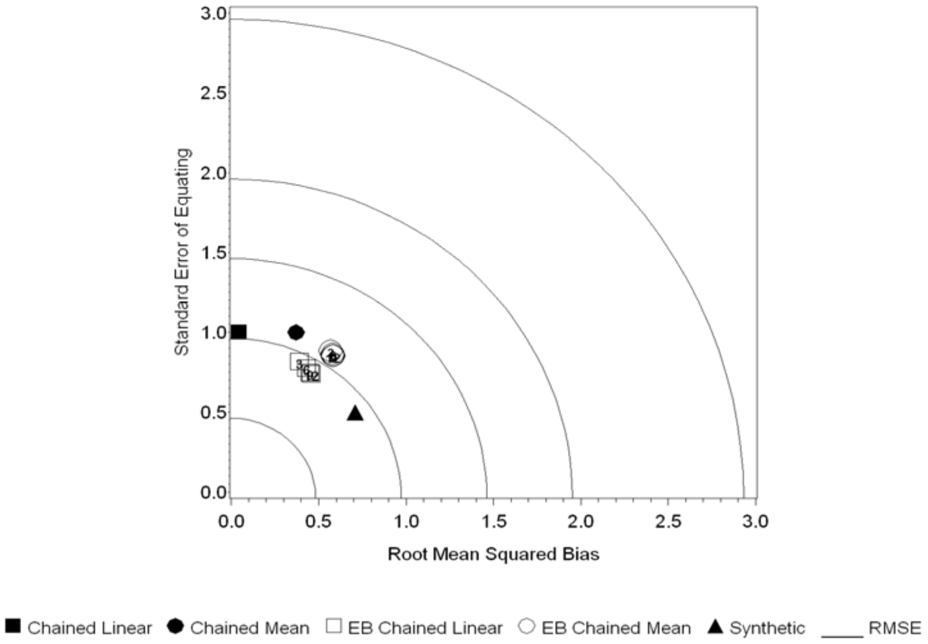


***Figure 11.* Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 75 in Study 1.**

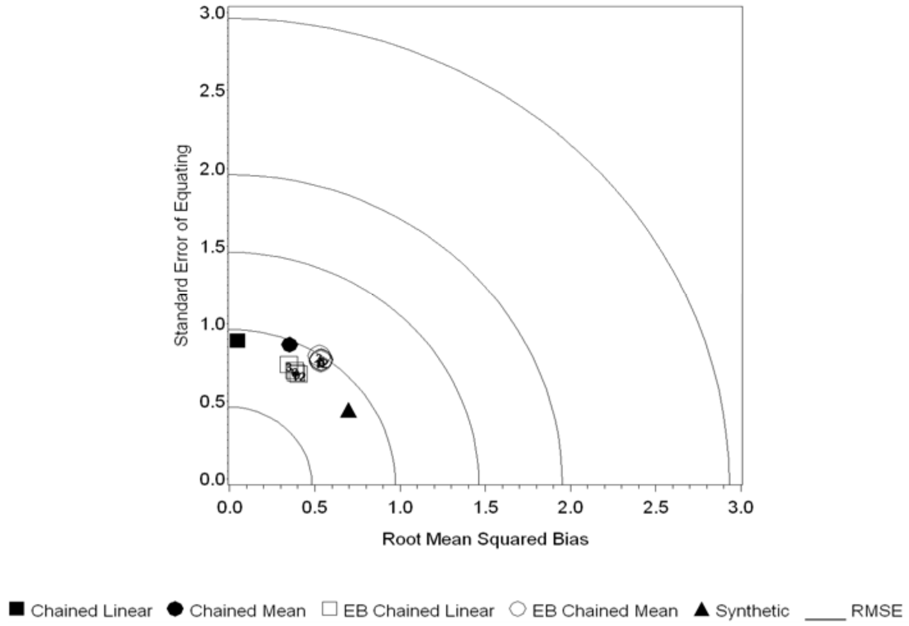


Figure 12. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 100 in Study 1.**

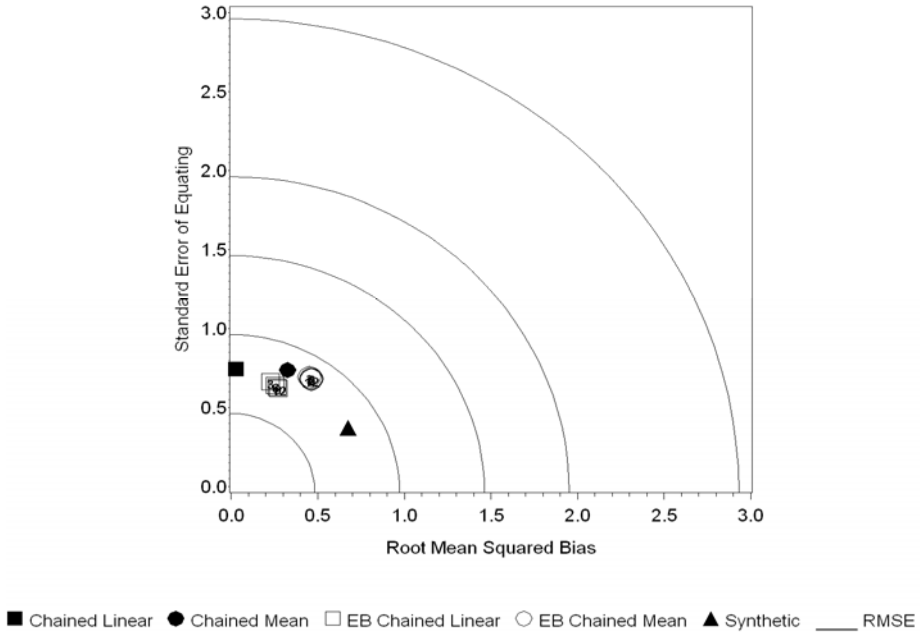


Figure 13. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 200 in Study 1.**

The average is a root-mean-square, so that negative bias at one score level cannot compensate for positive bias at another. The vertical position of the data point is the standard deviation (over the 500 replications) of the small-sample equated scores, averaged over score levels in the same way. The units of the horizontal and vertical scales are equal, so that the distance of each data point from the origin represents the weighted average RMSE over the 500 replications, averaged over score levels in the same way. The concentric arcs indicate selected values of the weighted average RMSE. In general, the four data points for an EB method appear on the graph in the order (3, 6, 9, 12), with the smallest number of prior equatings (3) resulting in the largest SEE and the largest RMSE. Table A1 in the appendix presents the summary of the weighted average root mean squared bias, SEE, and RMSE within each combination of sample size and equating method.

Figure 8 shows the results for equating with new-form samples of only 10 examinees. The small-sample equatings produced by the two EB methods showed a greater equating bias, but a much smaller SEE, which resulted in a substantially smaller RMSE than the non-EB methods did. Increasing the number of prior equatings reduced the SEE, while producing only a tiny increase in the bias, resulting in a smaller RMSE. Although the chained mean equating (indicated by the black circle) had a smaller SEE than did the chained linear equating (the black square) when no collateral information was used, the use of collateral information reversed this comparison; the EB chained linear method had a smaller SEE than did the EB chained mean method. The synthetic function created by averaging the chained linear equating with the identity performed about as well as the EB chained linear method that used all 12 prior equatings; its average equating error was smaller, and even its bias was smaller. Because of the way this synthetic function is defined, its SEE must be exactly half the SEE of the non-EB chained linear equating.

Figure 9 shows the results for equating with new-form samples of 25 examinees. Again, the EB methods showed a larger bias but a smaller SEE than their non-EB counterparts, leading to a smaller RMSE; slightly smaller for chained mean equating, substantially smaller for chained linear equating. The synthetic function had an average bias nearly as large as the EB chained linear methods, but a slightly smaller SEE, producing a smaller RMSE.

Figure 10 shows the results for equating with new-form samples of 50 examinees. The advantage of using the prior information was much smaller in this case. There was almost no advantage from using collateral information in chained mean equating, but in chained linear equating, there was approximately a 20% reduction in the RMSE. Again, the synthetic function had the smallest RMSE, just slightly smaller than that of the EB chained linear method. Figure 11 shows the results for equating with new-form samples of 75 examinees. The comparisons between methods are similar to those for new-form samples of 50 examinees.

Figure 12 shows the results for equating with new-form samples of 100 examinees. The RMSE for the EB chained linear method (.78) was the smallest, slightly smaller than for the synthetic function (.83). The RMSE for chained mean equating was about the same with and without prior information, and was slightly larger than that for the other methods. For samples of 200 examinees, the comparisons between methods are similar to those for new-form samples of 100 examinees. As shown in Figure 13, the RMSE for the EB chained linear method (.68) was the smallest, slightly smaller than for chained linear equating without prior information (.74).

Figure 14 shows how the overall RMSE of the EB chained linear method varied with the number of examinees in the new-form sample and the number of equatings used as collateral information. The smaller the sample, the greater the importance of the collateral information. With samples of only 10 examinees, including a larger number of equatings as collateral information produced a noticeable improvement in accuracy. With samples of 100 or more examinees, the improvement was tiny.

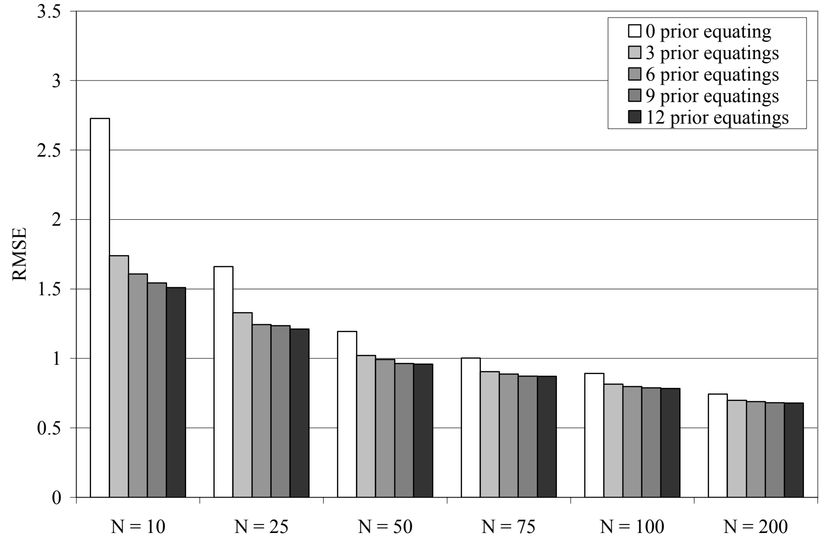


Figure 14. **Root mean squared error as a function of sample sizes and the number of priors in Study 1.**

Study 2

Data

The data sets used in Study 2 were drawn from the same test as those in Study 1. The April 2004 administration, which was the new-form administration in Study 1, became the reference-form administration (Population *Q*, Form *Y*) in Study 2, but with one additional item excluded from the total score. The March 2005 administration was the new-form administration (Population *P*, Form *X*) in Study 2. Forms *X* and *Y* each consisted of 107 MC items, of which 42 were anchor items. The test booklet for each form actually contained 110 items, but three items in each form were excluded from scoring. Descriptive statistics for Populations *P* and *Q* are summarized in Table 2.[8]

Table 2

Descriptive Statistics for the Observed Distributions of X, V in P and Y, V in Q: Study 2

	N	μ	σ	SEM	Reliability	ρ
X	6,426	73.62	10.51	4.2	.82	.91
V_P		30.60	4.96	2.3	.72	
Y	6,489	77.47	10.83	4.2	.84	.92
V_Q		30.46	5.09	2.3	.67	

Note. SEM = standard error of measurement, ρ = correlation between total score and anchor.

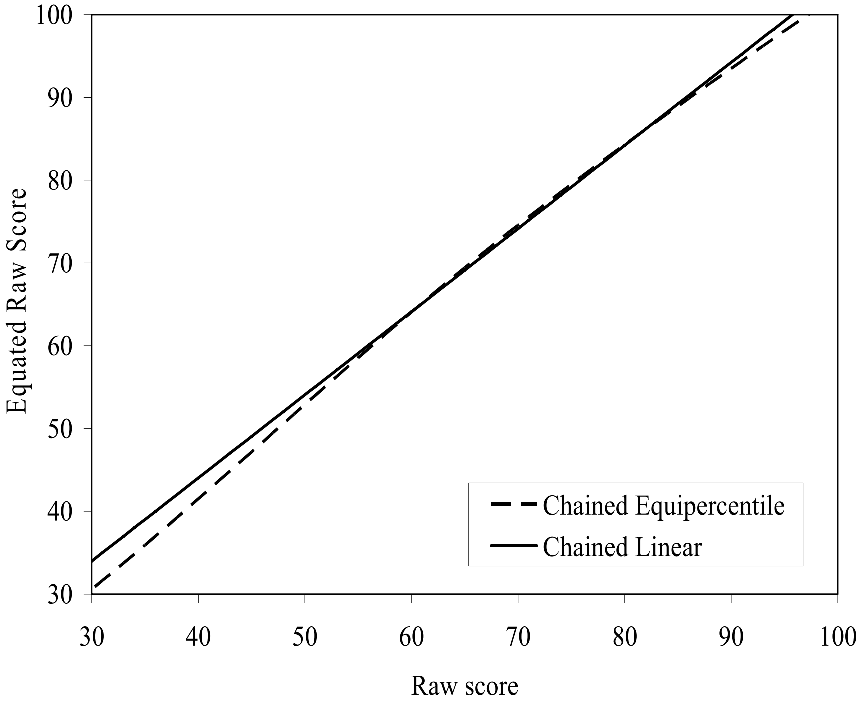
As Table 2 shows, population *P* was as adept as population *Q*. Their mean scores on the anchor differed by only 0.14 correct answers, an effect size of 0.03, indicating a negligible difference between the two populations. Therefore, the difference in the mean scores on the two forms (an effect size of -0.36) appears to be mainly a difference in the difficulty of the two forms. The difference in difficulty between the two forms to be equated was much greater in Study 2 than in Study 1. Otherwise, the two studies were essentially the same. In particular, the prior equatings used as collateral information in Study 2 were (with one exception) the same equatings used as collateral information in Study 1. The internal consistency reliabilities for both tests (.82 and .84) and anchors (.67 and .72) were moderate. The correlations between the tests and the internal anchors were high (.91 and .92).

The criterion equating was conducted with a total of 6,426 examinees for form *X* and 6,489 examinees for form *Y*. Figure 15 plots the raw to equated-raw score conversions produced by the chained linear and chained equipercentile methods, in the portion of the score range where any examinees' scores were observed. The differences between the two equating functions were substantial, at score levels that included many examinees. The criterion equating chosen was the raw-to-equated-raw score conversion derived from the chained equipercentile equating, which was curvilinear enough that it could not be approximated well by a linear equating function. This curvilinearity in the criterion equating limited the extent to which any of the small-sample equatings, which were all done by linear methods, could accurately reproduce the criterion equating.

Procedure

Study 2 was the same as Study 1, with respect to the resampling procedure, sample sizes, selection of collateral information, equating methods, and deviance measures. The pool of prior equatings was also the same, except for the addition of one more equating; the equating that provided the large-group data for Study 1. Consequently, a total of 13 prior equatings were available in Study 2. The numbers of prior equatings used as collateral information in the small-sample equatings were the same as in Study 1: 3, 6, 9, and 12.

Note that in Study 2, although the criterion function was nonlinear, the small-sample equating methods were linear (chained linear and chained mean). In practical equating situations, equating with small samples is commonly done by linear methods, because the data are not considered sufficient to estimate more than the means and standard deviations of the score distributions. If the actual (unknown) equating relationship, in the population of interest, is not linear, linear equating methods will be biased. Typically, they will be biased in one direction in the middle of the score distribution and in the opposite direction at the ends. In our EB procedure, the EB equating function is a compromise between the small-sample equating function (which is linear) and prior equatings (which may be either linear or nonlinear). Our EB procedure estimates the equated score separately for each possible raw score on the targeted new form, thus it can estimate nonlinear equating functions without assuming that they have a particular mathematical form. Therefore, both the EB chained linear and EB chained mean methods can produce a nonlinear equating transformation, even though the non-EB versions of these methods cannot. It is interesting to compare the EB and non-EB results, particularly where the large-group equipercentile equating differed substantially from the large-group linear equating.



Raw score	Frequency
0-30	1
31-36	6
37-42	15
43-48	41
49-54	114
55-60	224
61-66	438
67-72	855
73-78	1236
79-84	1418
85-90	1277
91-96	669
97-102	130
103-108	1
Total	6426

***Figure 15*. Plot of raw-to-equated-raw score for chained linear and chained equipercentile, and frequency distribution of new form *X* scores in total group: Study 2.**

Results

Figures 16 to 21 show plots of the conditional RMSE for each method with each new-form sample size. Figures A7 to A12 in the appendix show the conditional bias and CSEE for the six new-form sample sizes. As in Study 1, the EB results shown in these figures are from the small-sample equatings that used 12 prior equatings as collateral information.

In general, no single method dominated the others. With new-form samples of only 10 examinees, as shown in Figure 16, chained mean equating performed fairly well throughout most of the score range. Adding collateral information made it perform better for very low and very high scores, but more poorly from about -1.5 SD to about +1 SD. Chained linear equating performed well for scores within about 1 SD of the mean, but very poorly for scores more than 1.5 SD from the mean. Adding the collateral information improved its performance for these low and high scores, but degraded its performance between -1.5 and +1.5 SD. The synthetic function's performance was similar to that of (non-EB) chained mean equating; somewhat less accurate than chained mean equating for scores between -2 SD and +1 SD, but more accurate for scores above +1 SD.

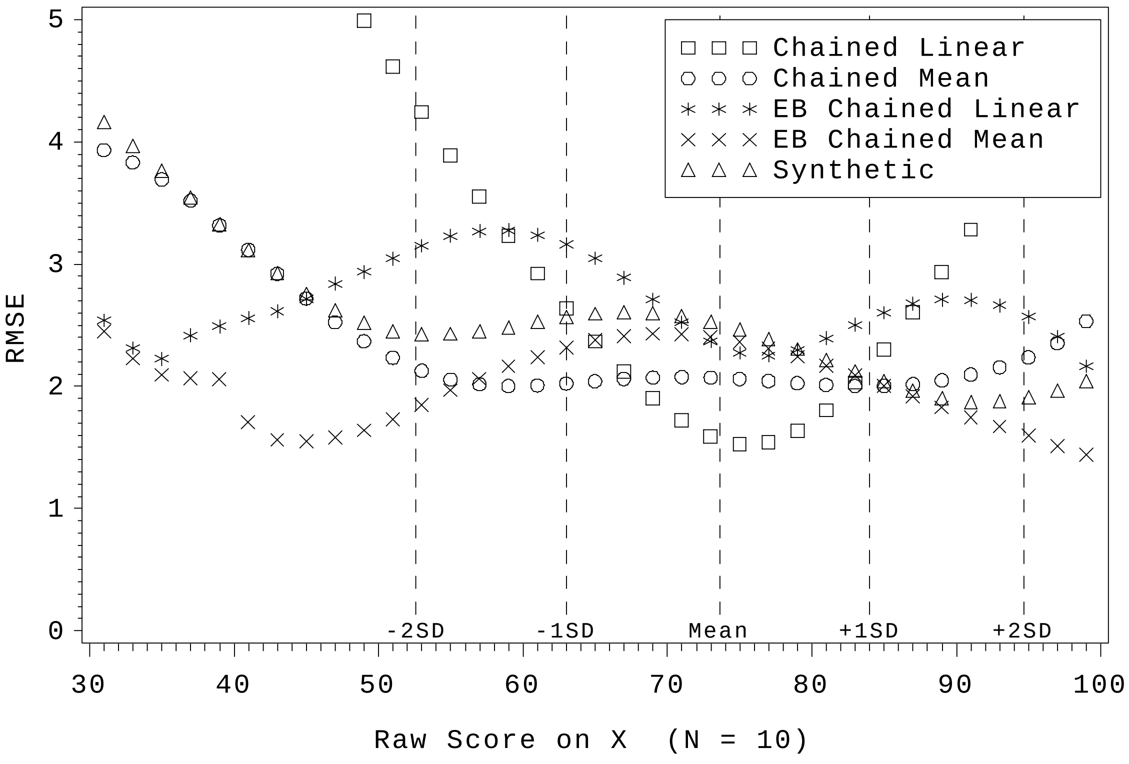


Figure 16. **Conditional root mean squared error at samples of 10 in Study 2.**

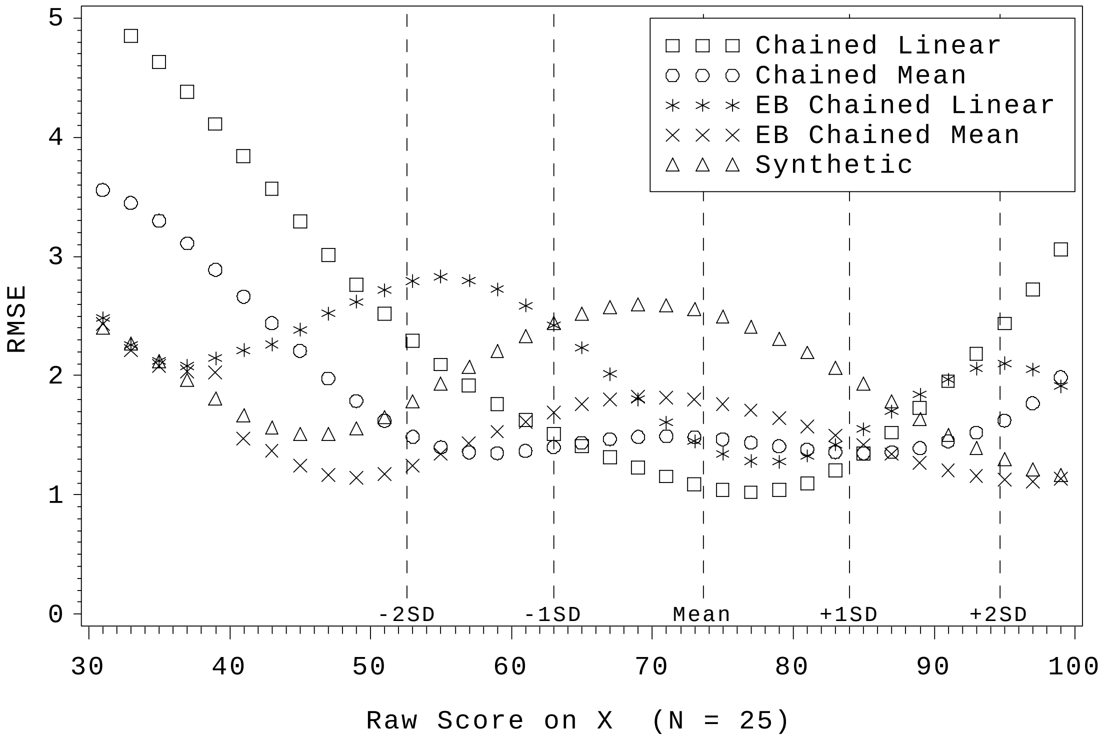


Figure 17. **Conditional root mean squared error at samples of 25 in Study 2.**

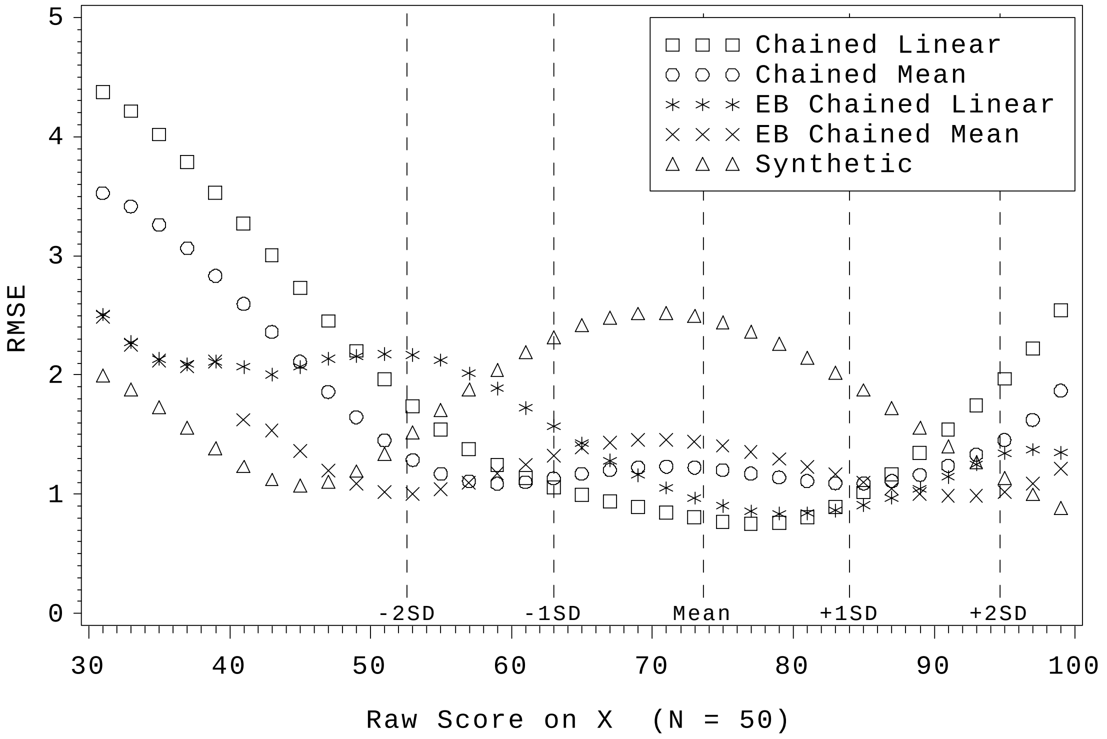


Figure 18. **Conditional root mean squared error at samples of 50 in Study 2.**

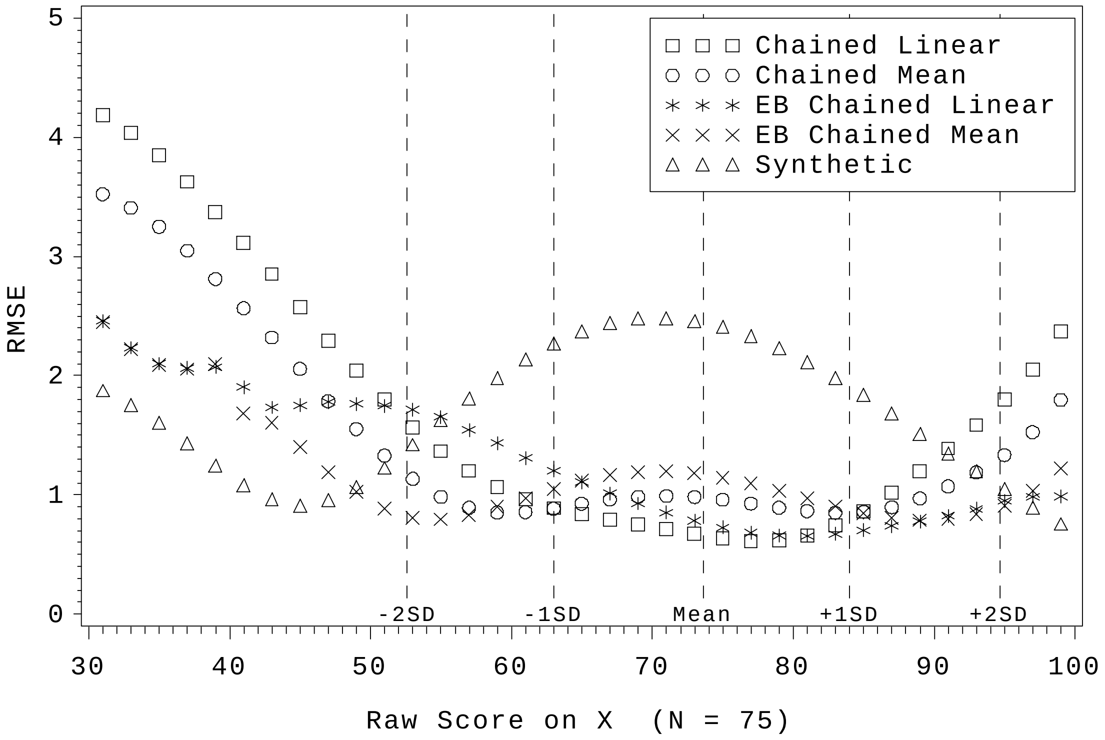


Figure 19. **Conditional root mean squared error at samples of 75 in Study 2.**

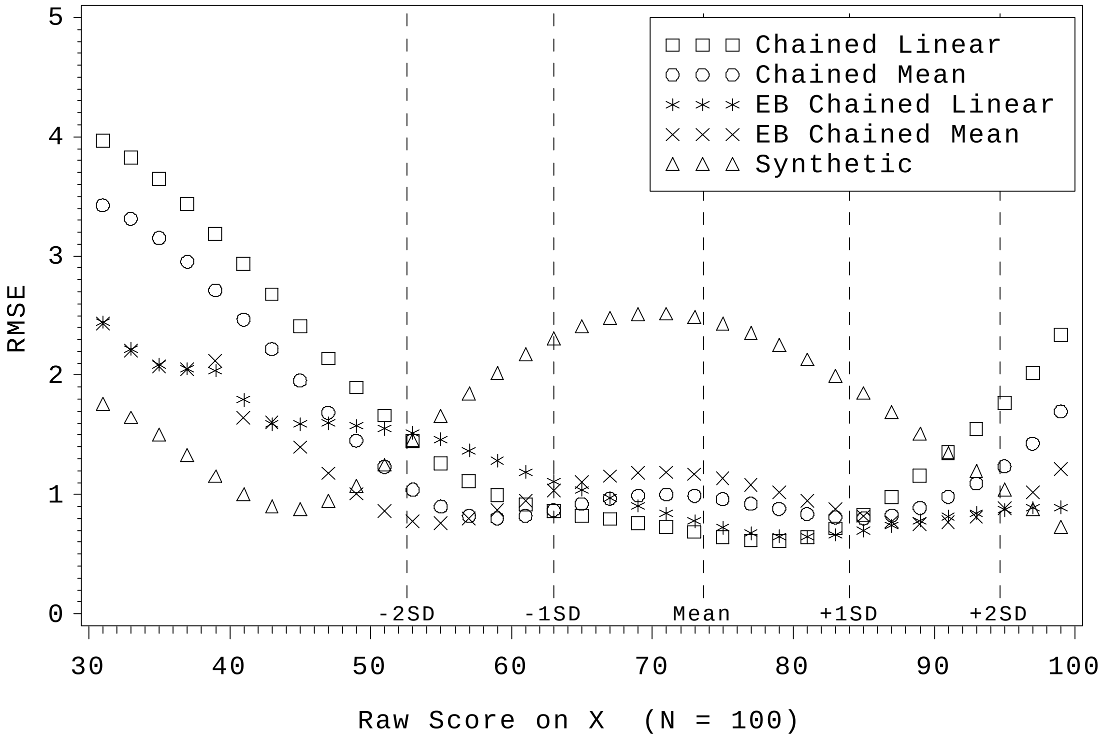


Figure 20. **Conditional root mean squared error at samples of 100 in Study 2.**

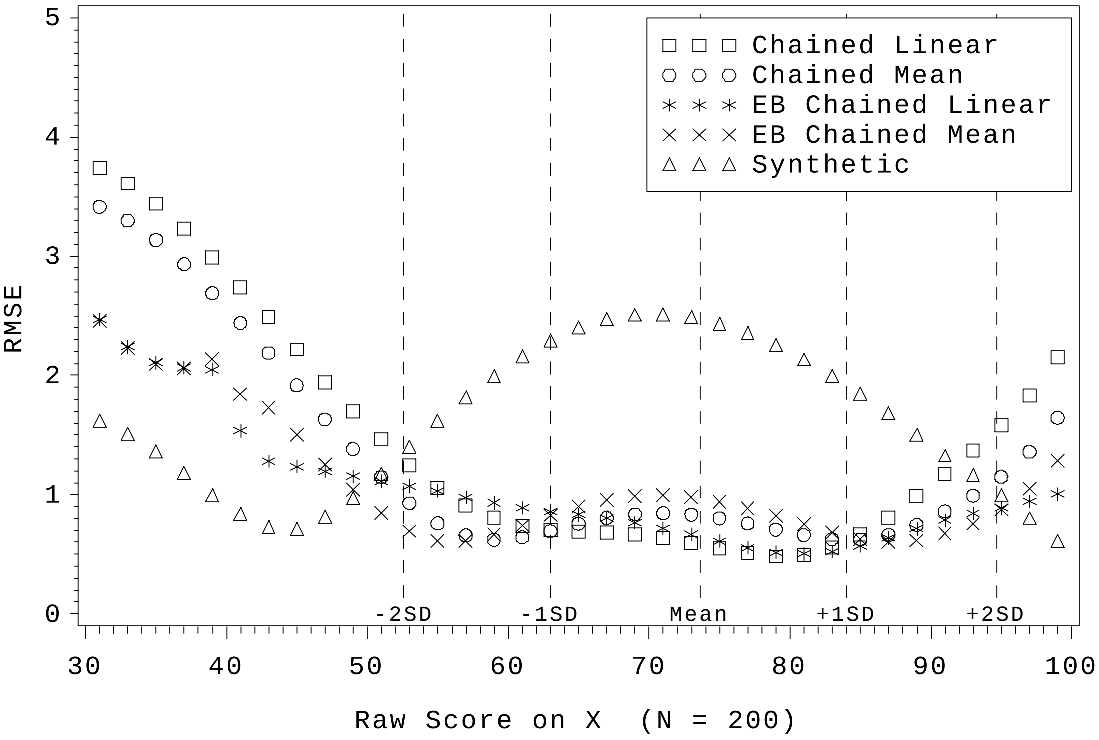


Figure 21. **Conditional root mean squared error at samples of 200 in Study 2.**

As the new-form sample size increased, the two non-EB methods became more accurate. The chained linear equating continued to produce the most accurate results for scores less than 1 SD from the mean and the least accurate results for scores more than 2 SD from the mean. The two EB methods became more similar to the corresponding non-EB methods, and therefore more accurate. Unlike the other methods, the synthetic function did not improve noticeably in accuracy as the new-form sample size increases, so it became the least accurate of the methods compared.

Figures 22 to 27 show the RMSE decomposed into its two orthogonal components of bias and standard error and averaged over score levels, weighting by the population frequency, as in Study 1. Table A2 in the appendix presents the summary of those deviance measures. For all methods, as the new-form sample size increased, the standard error decreased. The bias decreased for the two EB methods, because the larger sample size resulted in a smaller weight being given to the collateral information from the prior equatings. In general, the standard error of each method in Study 2 was somewhat smaller than in Study 1, but the bias was much larger.

The larger bias observed here for the chained linear and chained mean equating methods is possibly due to the nonlinearity of the criterion equating.[9] The bias was especially large for the synthetic function that included the identity, because the new form and the reference form differed substantially in difficulty. Because the weight given to the identity did not decrease with increasing sample size, the bias in the results of the synthetic function did not decrease.

Figure 28 shows how the overall RMSE of the EB chained linear method varied with the number of examinees in the new-form sample and the number of equatings used as prior information. As in Study 1 when the sample was small, collateral information based on more prior equatings had a greater effect, but in Study 2, that effect made the EB equating less accurate. When the new-form sample was larger, 100 or 200, however, more collateral information tended to make the equating slightly more accurate.

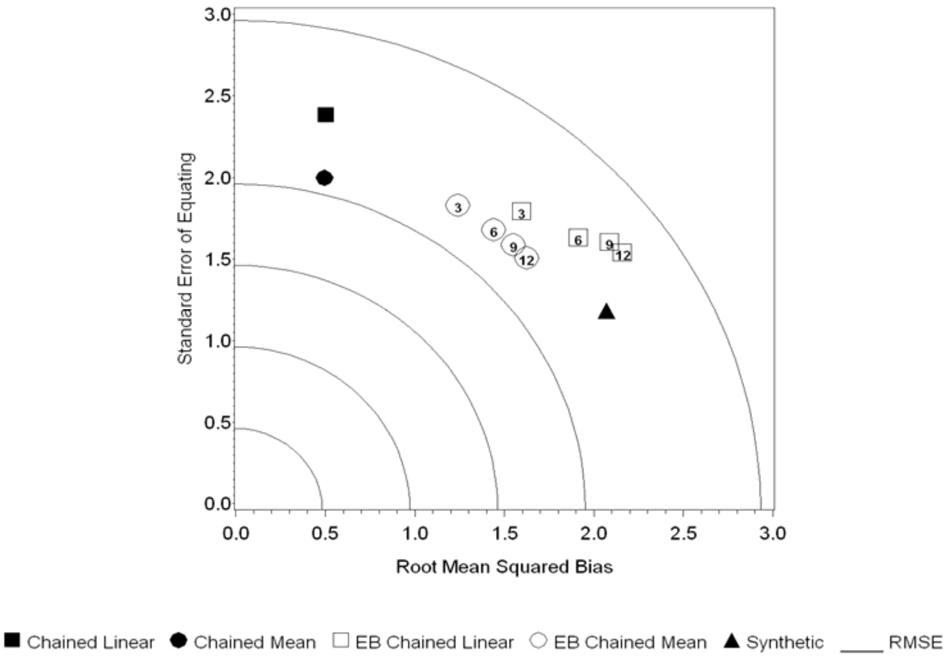


***Figure 22.* Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 10 in Study 2.**

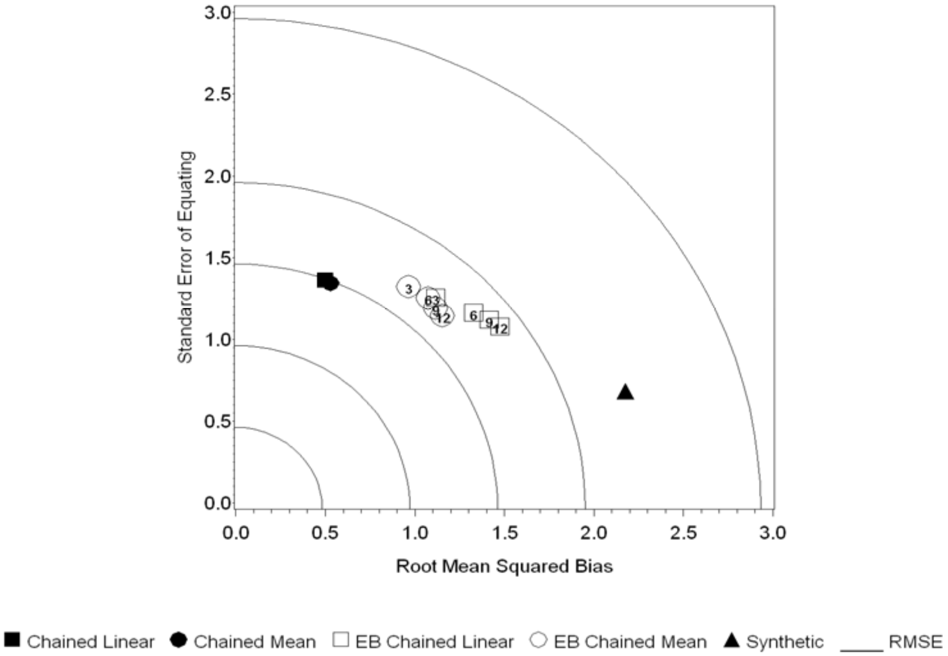


Figure 23. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 25 in Study 2.**

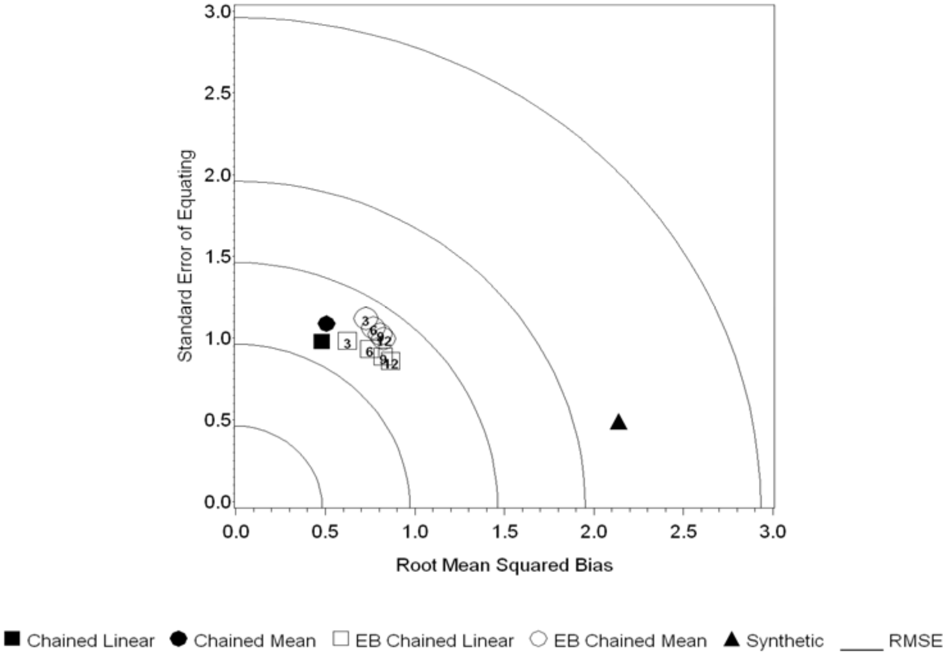


Figure 24. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 50 in Study 2.**

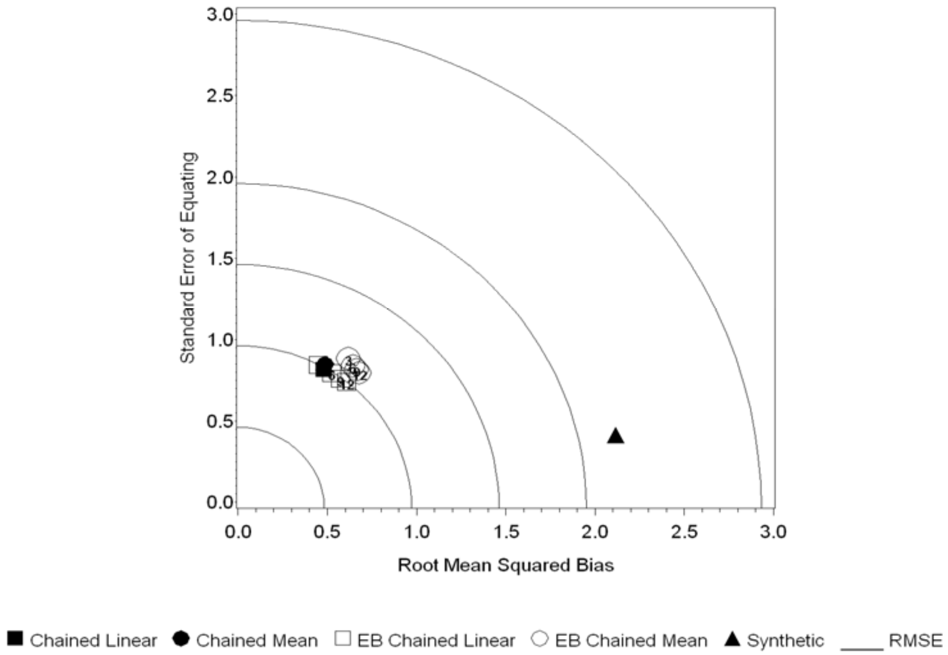


Figure 25. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 75 in Study 2.**

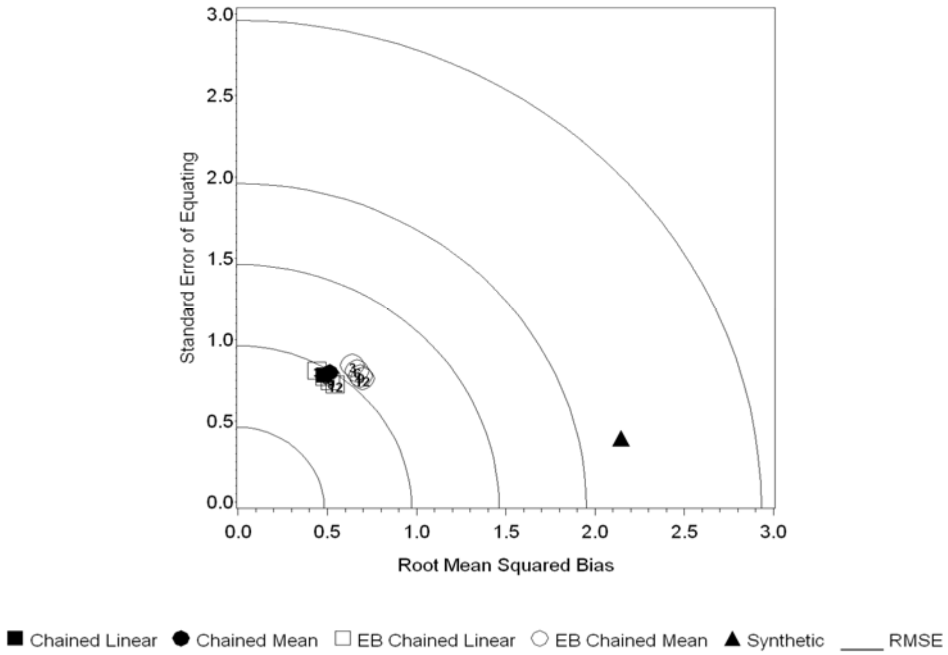


Figure 26. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 100 in Study 2.**

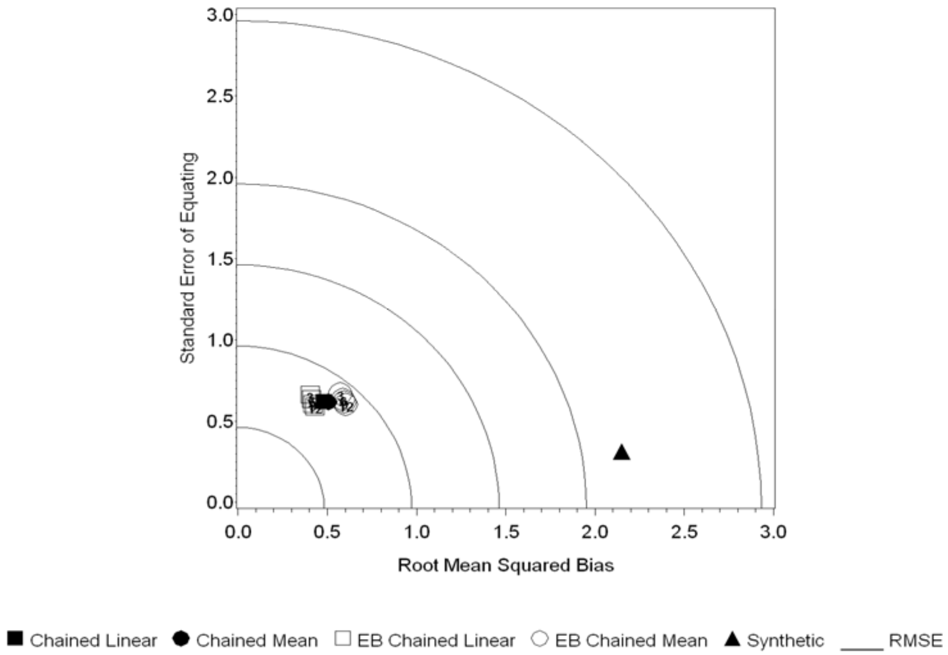


Figure 27. **Plot of the weighted average root mean squared error as a function of the weighted average root mean squared bias and error at samples of 200 in Study 2.**

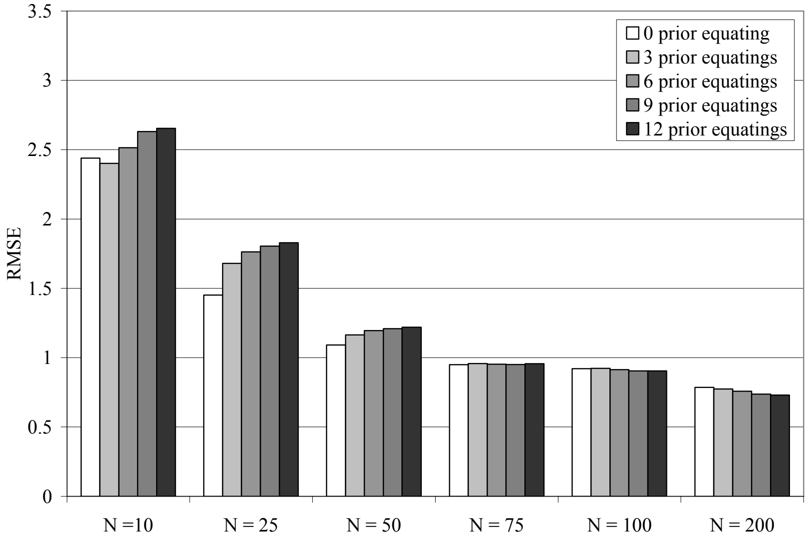


Figure 28. **Root mean squared error as a function of sample sizes and the number of priors in Study 2.**

Discussion

In these resampling studies, an EB procedure was evaluated for its effect in improving the accuracy of small-sample equating by incorporating collateral information from the equating of other forms of the same test. The EB procedure evaluated in these studies estimates equated each score separately (point-by-point estimation), rather than estimating the parameters of the equating function (e.g., slope and intercept). This procedure allows the estimation of nonlinear equating functions and the use of collateral information from test forms having different numbers of items.

Although Studies 1 and 2 involved the equating of two different forms of the same test, using nearly the same collateral information, the results of the two studies were quite different. In Study 1, the collateral information improved the accuracy of the equating, particularly when the new-form sample was very small. In Study 2, it did not. A synthetic function formed from the (non-EB) chained linear equating and the identity, weighted equally, produced the most accurate results in Study 1 and the least accurate results in Study 2. The reason for the difference was that in Study 1 the new form and reference form were similar in difficulty, as they were in the set of prior equatings used as collateral information. Consequently, both the prior equatings (in the EB methods) and the identity (in the synthetic function) tended to pull the small-sample equating results toward the correct equating transformation. In Study 2, the new form was considerably more difficult than the reference form. Consequently, both the prior equatings (in the EB methods) and the identity (in the synthetic function) tended to pull the small-sample equating results away from the correct equating transformation. In the EB equatings, the influence of the prior equatings decreased as the size of the new-form sample increased. In the synthetic function, it did not, because the weight given to the current equating was constant (0.5 in this case), regardless of the sample size or the standard error.

Overall, using collateral information rather than conventional equating alone seems promising for equating with very small samples. The results of the present study were somewhat inconsistent, though, leading to different conclusions regarding the use of the EB method with small samples. As Study 2 indicated, the EB method may not always be a good choice. The EB method was derived from the idea that any increase in systematic error a prior produces is more than offset by the decrease in random error. As expected, a common trend that our studies confirmed was that the EB method yielded greater bias, but less equating error, than did traditional linear equating methods. The RMSE index indicated, however, that the trade-off between bias and error with EB methods seemed to depend on the similarity in difficulty between the targeted new form and previous equating forms. The EB method is based upon the fundamental assumption that the pair of forms in the current equating is sampled from the same domain as the pairs of forms in the prior equatings. In Study 1, the targeted new form was similar in difficulty to the average of other forms in the same domain, thus, incorporating collateral information improved equating accuracy with small samples, by reducing sampling error. In Study 2, the targeted new form was sampled from the same domain as the prior forms, but it appeared to be an outlier in the domain, with respect to difficulty. Consequently, the use of collateral information was disadvantageous, due to the substantial bias produced. The EB method's effectiveness depends heavily on the selection of collateral information, a demanding job, in practice.

One possible objection to the EB procedure used in these studies is that it is not truly an equating method, because it is not symmetric. That is, interchanging the new form and the reference form will not produce the exact inverse of the equating transformation. To be entirely correct, this EB procedure should properly be described not as an equating method, but as a procedure for estimating an equating transformation. The lack of symmetry might possibly introduce bias into the procedure, but the amount of bias will be negligible, especially in relation to the instability of any estimate (symmetric or not) based on small samples.

Previous studies (Kim et al., 2006, 2007) addressed the benefits of averaging the identity with conventional equating functions when equating samples are small. Like the EB approach, the synthetic function is an average of the current equating and a prior estimate of the population equating. Instead of determining the prior estimate from previous equatings, it uses the identity; a logical choice, if there is no way to know in advance whether the new form is more likely to be easier or harder than the reference form. The major problem with the synthetic function used in these studies is that the weight assigned to the current equating does not depend on the size of the samples. The synthetic function would perform better if the weights were chosen to give the current equating greater weight when it is based on larger samples of examinees. Modifications are needed to update the weight for the current equating (and the identity as a prior) on the basis of the available data, as in the EB estimation procedure. The synthetic function on this case can be called the Bayesian synthetic function. Comparison of the EB method with the Bayesian synthetic function is a good topic for future study.

In the present study, all prior equatings were chosen from the same test as the targeted new forms, and they were based on fairly large samples, ranging from 300 to 6,000 examinees. It would be very difficult to find many prior equatings derived from large sample data, however, particularly for low volume tests. The use of collateral information is the preferred method in situations where equating must be conducted with little data available. Because, in practice, only a few forms are available for low-volume tests, the practical implication of the EB method would be extremely limited, under the assumption that prior equatings should come from the same test as the targeted new test form. We think that collateral information could also include equatings of other tests having similar content structures or specifications. To support this argument, however, substantial empirical evidence should be displayed, based on various real data sets. The EB study using collateral information from different sources and various sample sizes will be presented elsewhere in detail.

The present study showed that the effectiveness of the EB estimation procedure depends heavily on the selection of collateral information. When using collateral information from different tests to enhance the accuracy of small-volume equating, gathering prior equatings from other tests would be more demanding than gathering them from the same test as the targeted new test form. This selection procedure seems to be somewhat subjective; but this subjectivity should be differentiated from an arbitrary or blind decision rule. Psychometricians need to exercise their best judgment (subjective) when gathering collateral information and base their decisions upon several test characteristics such as test length, equating design and method, the length of anchor, and item format (e.g., set items).

The purpose of equating is to provide examinees with fair and accurate scores, by adjusting for differences in form difficulty. Data collection is the most important factor in test equating (Holland, Dorans, & Petersen, 2007), and the sample size problem is essentially a data collection problem. As the literature indicates, equating with small samples is very difficult, because none of the equating methods are expected to perform properly. Several methodological approaches (e.g., smoothing, creation of a synthetic function) have been proposed, to improve the practice of equating with little data. In this study, the effectiveness of using collateral information was assessed as an alternative for equating with little data. Dorans (personal communication, November 9, 2007) emphasized that no methodology can solve problems caused by a lack of data. Given that problem, reporting raw scores rather than scaled scores might be considered as an alternative to the extremely small sample (e.g., fewer than 30) problem until adequate data become available (N. J. Dorans, personal communication, November 9, 2007). Nevertheless, many practitioners must use little data to make the best estimates of the equating transformation in operational testing situations, because many testing programs need to report comparable scores for a new edition of an established test form in a timely manner, regardless of sample sizes. Thus, along with improving data collection procedure, it is important to seek methods that have the potential to address the problems involved in small-sample equating.

References

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: Academic.

Dorans, N. J., & Feigenbaum, M .D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM-94-10; pp. 91-122). Princeton, NJ: ETS.

Harris, D. J. (1993, April). *Practical issues in equating.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 169-203). Amsterdam: Elsevier B. V.

Hsu, T., Wu, K., Yu, J. W., & Lee, M. (2002). Exploring the feasibility of collateral information test equating. *International Journal of Testing, 2,* 1-14.

Kim, S., von Davier, A. A., & Haberman, S. (2006, April). *An alternative to equating with small samples in the non-equivalent groups anchor test design*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

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.

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

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational measurement, 19*, 279-293.

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

Livingston, S. A., & Lewis, C. (2007). *Small sample equating with prior information*. Unpublished manuscript.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30*, 55-78.

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

Sinharay, S., Dorans, N. J., Grant, M. C., Blew, E. O., & Knorr, C. M. (2006). *Using past data to enhance small-sample DIF estimation: A Bayesian approach*. Unpublished manuscript.

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

Notes

[1] In a single level Bayesian analysis, the data on which priors are set needs to be independent of data on the form from which the parameter estimates are generated. The mean and variance of priors in a population is based solely on old data in the single level Bayesian analysis. In the hierarchical Bayesian framework, however, both prior and current equating data are regarded as the first-level parameters, and they are conditioned on the second-level parameters (e.g., the mean and variance of true values for equatings in a population), which indicate the mean and variance of prior equatings in this study. Therefore, the current equating should be utilized along with the old equating data when estimating the mean and variance of the prior in the population.

[2] Different forms of the same test sometimes have different numbers of items, either by design or, more often, because of the removal of problematic items before the scores are calculated.

[3] Only linear equating functions were considered, because the sample of interest was too small to ensure the adequacy of equipercentile equating results (Harris, 1993; Kolen & Brennan, 2004).

[4] The mean equating function used in this study is slightly different from the mean equating function presented in Kolen and Brennan's book (2004, p. 125). Our version of mean equating follows the same logic as chained linear equating. For that reason, we called it chained mean equating.

[5] We computed the weighted root mean squared bias to prevent large positive and negative differences from cancelling each other out.

[6] The test booklet for each form actually contained 110 items, but two items in each form were excluded from scoring.

[7] These numbers do not include the current small-sample equating, which was included in the calculation of the prior estimate.

[8] The descriptive statistics for Population Q in Study 2 differ slightly from those for Population P in Study 1, because of the exclusion of one additional item from the scoring in Study 2 and the inclusion in Study 2 of some examinees whom were excluded from Study 1.

[9] We also conducted equating analysis using the chained linear equating function as the criterion. In those analyses, the bias of the non-EB linear methods was much smaller. Although all the face values were different, the general trends for the EB, non-EB, and synthetic function methods were almost identical, regardless of the criterion function.

Appendix

The General Empirical Bayes Procedures

1. Convert raw scores on the targeted new form to percent-correct scores, ranging from 0 to 100.
2. Convert both raw and equated raw scores, along with the conditional standard errors of equating (CSEE), to the percent-correct metric for each prior equating.
3. Find the percent-correct raw score on each prior form that corresponds to each percent-correct score on the targeted new form.
4. Find the percent-correct equated raw score that corresponds to each percent-correct raw score on the prior form.
5. Find the percent-correct CSEE that corresponds to each percent-correct raw score on the prior form.
6. Use interpolation if the percent-correct score on the targeted new form is between two percent-correct values on the prior form.
7. Repeat these procedures for all prior equatings available.
8. Convert the matched percent-correct scores to the raw score unit of the targeted reference form by multiplying the percent-correct by the maximum number-correct score of the target reference form in each prior equating.
9. Find the equated raw score (y_{eq}) and squared CSEE (σ_{eq}^2) at each score point for the actual targeted new form equating.
10. For each raw score point, compute the mean (y_{prior}) and variance (σ_{prior}^2) across the corresponding equated raw scores for both the prior and targeted new forms. Adjust the variance by incorporating the squared CSEEs for all the forms using Equation 2.
11. Compute $\hat{y}_{EB}$ using Equation 1.

Table A1

Summary of Deviance Measures Across the Entire Score Region: Study 1

Sample size	CHLN	CHME	EB chained linear				EB chained mean				Syn.
			n of priors				*n* of priors				
			3	6	9	12	3	6	9	12	
					Root mean squared bias						
10	0.10	0.41	0.83	0.95	1.01	1.03	0.85	0.96	1.00	1.02	0.66
25	0.13	0.45	0.73	0.78	0.81	0.82	0.77	0.80	0.82	0.82	0.74
50	0.05	0.30	0.42	0.50	0.50	0.52	0.53	0.57	0.57	0.57	0.67
75	0.04	0.38	0.39	0.43	0.45	0.46	0.57	0.58	0.58	0.58	0.71
100	0.05	0.36	0.35	0.38	0.39	0.40	0.53	0.54	0.53	0.54	0.70
200	0.03	0.33	0.23	0.26	0.27	0.27	0.45	0.46	0.47	0.46	0.68
					Standard error of equating						
10	2.73	2.40	1.53	1.30	1.17	1.11	1.76	1.61	1.53	1.48	1.34
25	1.66	1.61	1.11	0.97	0.93	0.89	1.28	1.18	1.18	1.15	0.82
50	1.19	1.20	0.93	0.86	0.82	0.81	1.03	1.00	0.99	0.97	0.60
75	1.00	1.00	0.82	0.78	0.75	0.74	0.89	0.87	0.86	0.85	0.50
100	0.89	0.87	0.74	0.70	0.68	0.67	0.79	0.77	0.76	0.76	0.44
200	0.74	0.74	0.66	0.64	0.63	0.62	0.69	0.69	0.68	0.68	0.37
					Root mean squared error (RMSE)						
10	2.73	2.44	1.74	1.61	1.54	1.51	1.95	1.87	1.83	1.79	1.49
25	1.66	1.67	1.33	1.24	1.24	1.21	1.50	1.43	1.43	1.41	1.11
50	1.19	1.23	1.02	0.99	0.96	0.96	1.15	1.15	1.14	1.13	0.90
75	1.00	1.07	0.90	0.89	0.87	0.87	1.05	1.04	1.04	1.03	0.87
100	0.89	0.94	0.81	0.80	0.79	0.78	0.95	0.94	0.93	0.93	0.83
200	0.74	0.81	0.70	0.69	0.68	0.68	0.83	0.83	0.82	0.82	0.77

Note. CHLN = chained linear, CHME = chained mean, EB = empirical Bayes, Syn. = synthetic function.

Table A2

Summary of Deviance Measures Across the Entire Score Region: Study 2

Sample size	CHLN	CHME	EB chained linear *n* of priors				EB chained mean *n* of priors				Syn.
			3	6	9	12	3	6	9	12	
					Root mean squared bias						
10	0.50	0.50	1.60	1.91	2.09	2.16	1.24	1.44	1.55	1.62	2.07
25	0.50	0.54	1.12	1.33	1.42	1.48	0.97	1.07	1.12	1.15	2.18
50	0.48	0.51	0.62	0.75	0.82	0.87	0.73	0.77	0.81	0.83	2.14
75	0.48	0.49	0.45	0.52	0.57	0.61	0.62	0.64	0.66	0.68	2.12
100	0.48	0.52	0.44	0.49	0.52	0.54	0.64	0.67	0.69	0.70	2.15
200	0.48	0.51	0.41	0.41	0.42	0.43	0.57	0.59	0.59	0.60	2.15
					Standard error of equating						
10	2.39	2.00	1.79	1.63	1.60	1.54	1.83	1.68	1.59	1.51	1.18
25	1.36	1.35	1.26	1.16	1.12	1.08	1.32	1.25	1.19	1.15	0.68
50	0.98	1.09	0.98	0.93	0.89	0.86	1.12	1.06	1.02	1.00	0.49
75	0.82	0.85	0.85	0.80	0.76	0.74	0.89	0.84	0.82	0.80	0.41
100	0.78	0.80	0.81	0.77	0.74	0.72	0.84	0.81	0.78	0.76	0.39
200	0.62	0.62	0.66	0.64	0.61	0.59	0.67	0.64	0.62	0.61	0.31
					Root mean squared error (RMSE)						
10	2.44	2.07	2.40	2.52	2.63	2.65	2.21	2.21	2.22	2.22	2.39
25	1.45	1.45	1.68	1.76	1.81	1.83	1.64	1.65	1.64	1.63	2.28
50	1.09	1.20	1.16	1.20	1.21	1.22	1.34	1.31	1.30	1.29	2.20
75	0.95	0.98	0.96	0.95	0.95	0.96	1.08	1.06	1.05	1.05	2.16
100	0.92	0.95	0.92	0.91	0.91	0.91	1.06	1.05	1.04	1.03	2.18
200	0.78	0.80	0.78	0.76	0.74	0.73	0.88	0.87	0.86	0.86	2.18

Note. CHLN = chained linear, CHME = chained mean, *Note.* CHLN = chained linear, CHME = chained mean, EB = empirical Bayes, Syn. = synthetic function.

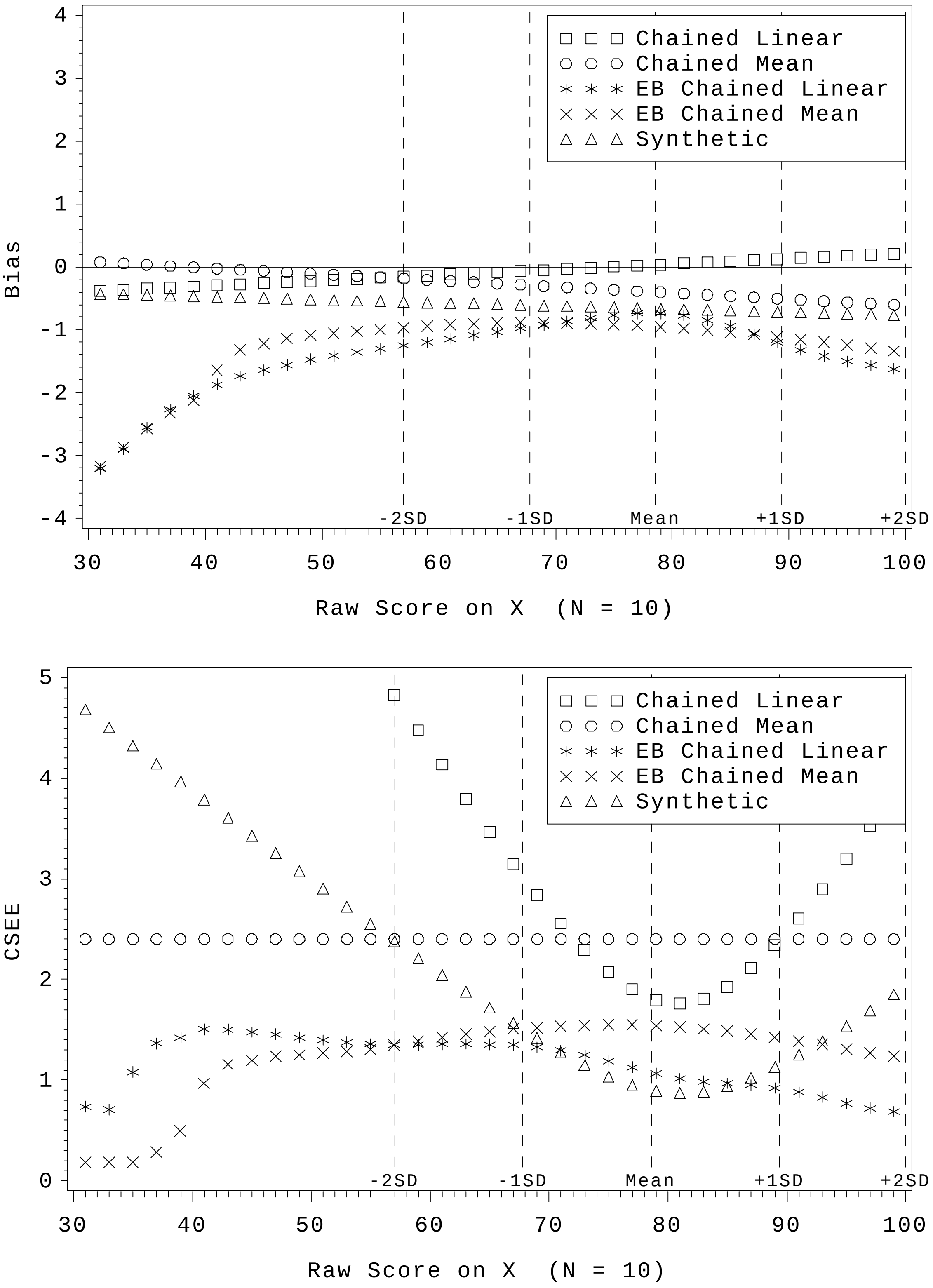


Figure A1. **Conditional bias and conditional standard error of equating (CSEE) at samples of 10 examinees in Study 1.**

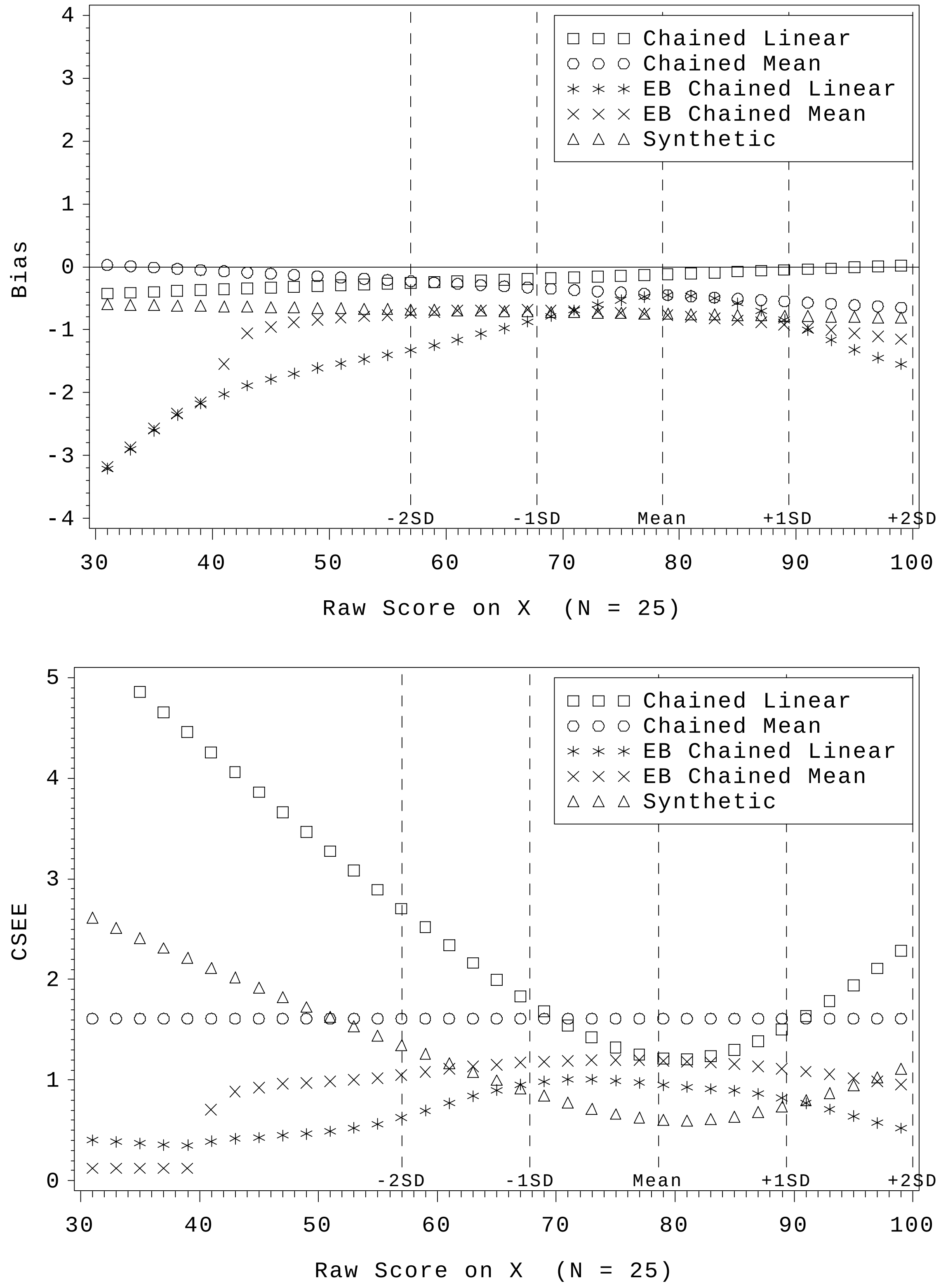
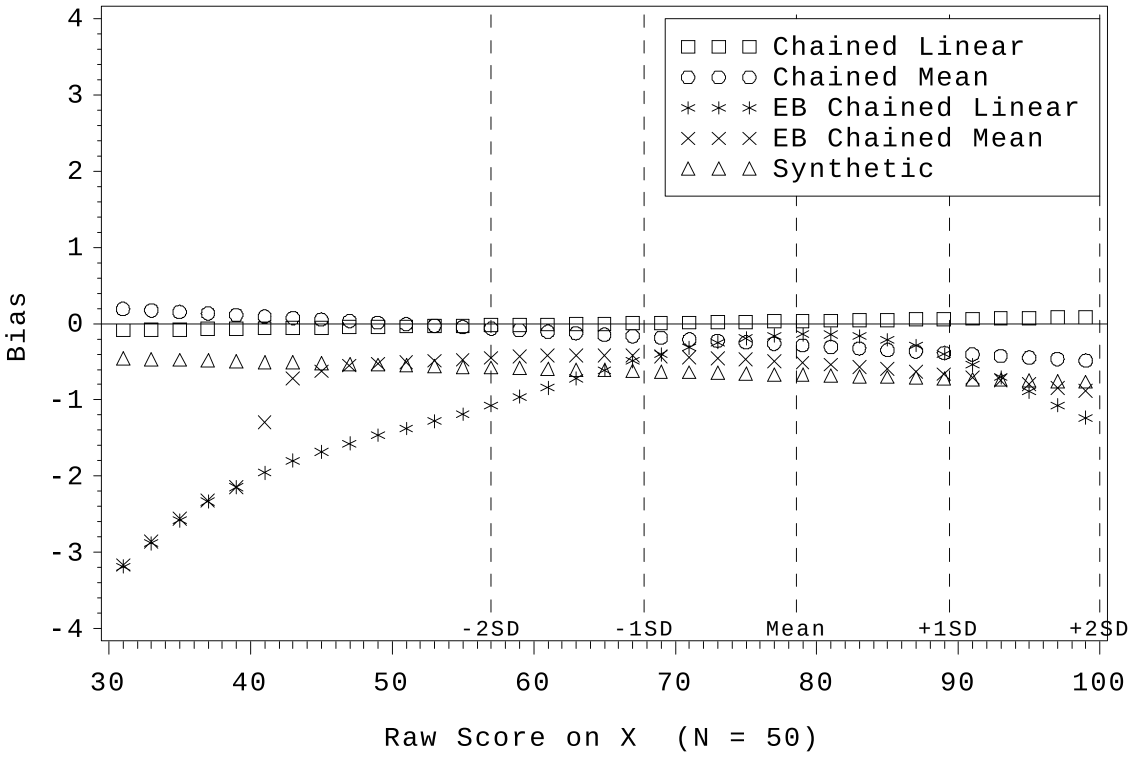


Figure A2. **Conditional bias and conditional standard error of equating (CSEE) at samples of 25 examinees in Study 1.**



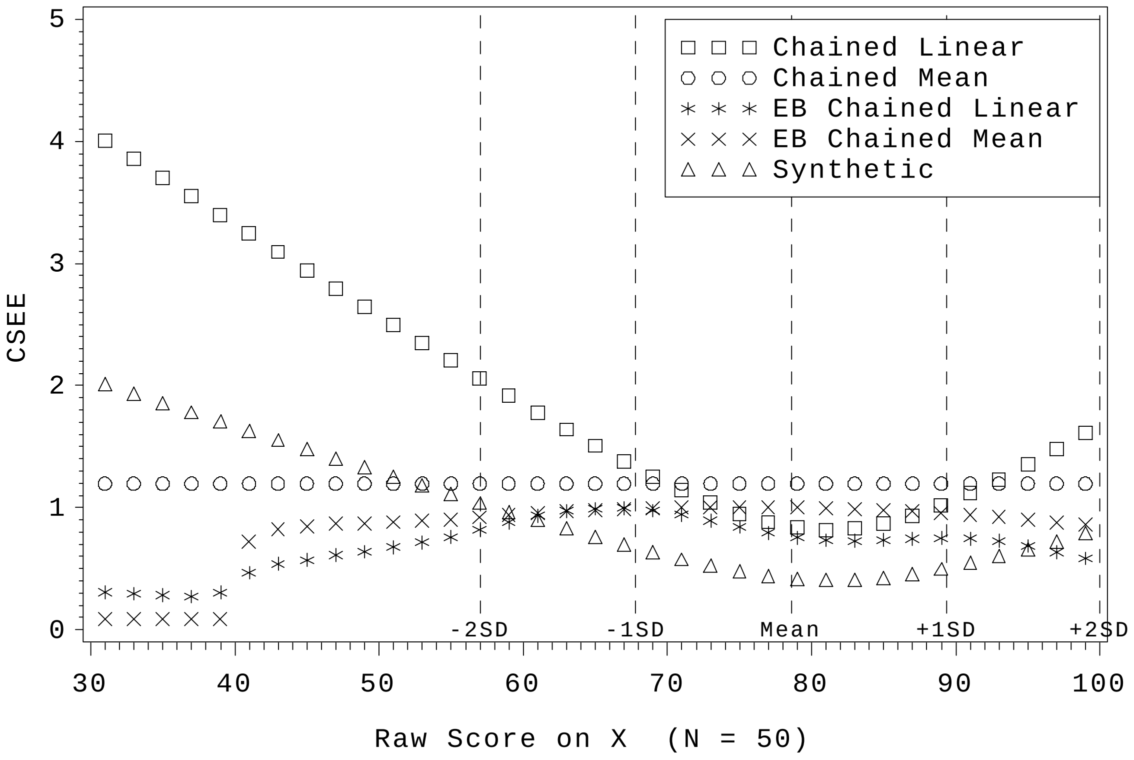


Figure A3. **Conditional bias and conditional standard error of equating (CSEE) at samples of 50 examinees in Study 1.**

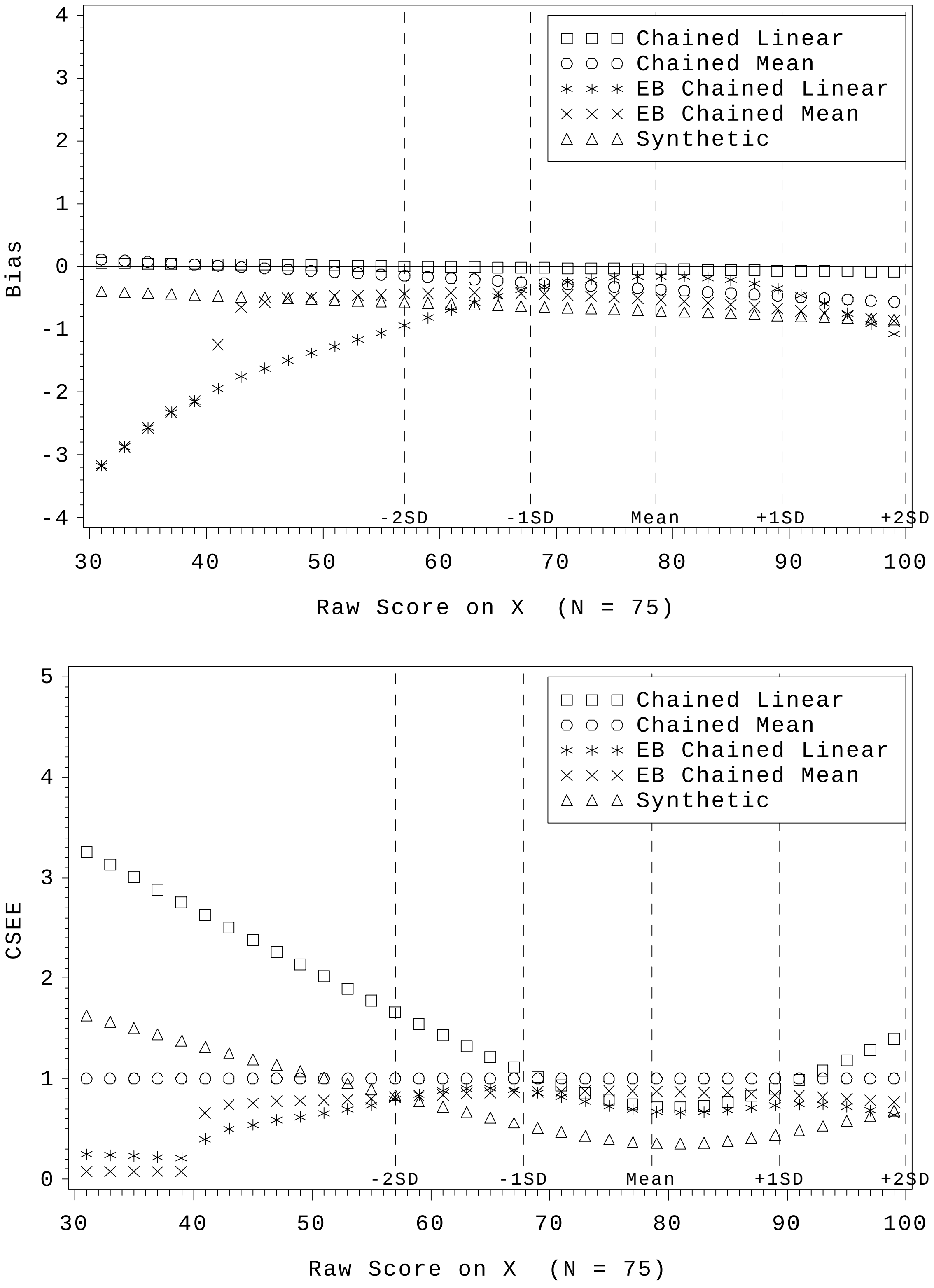


Figure A4. **Conditional bias and conditional standard error of equating (CSEE) at samples of 75 examinees in Study 1.**

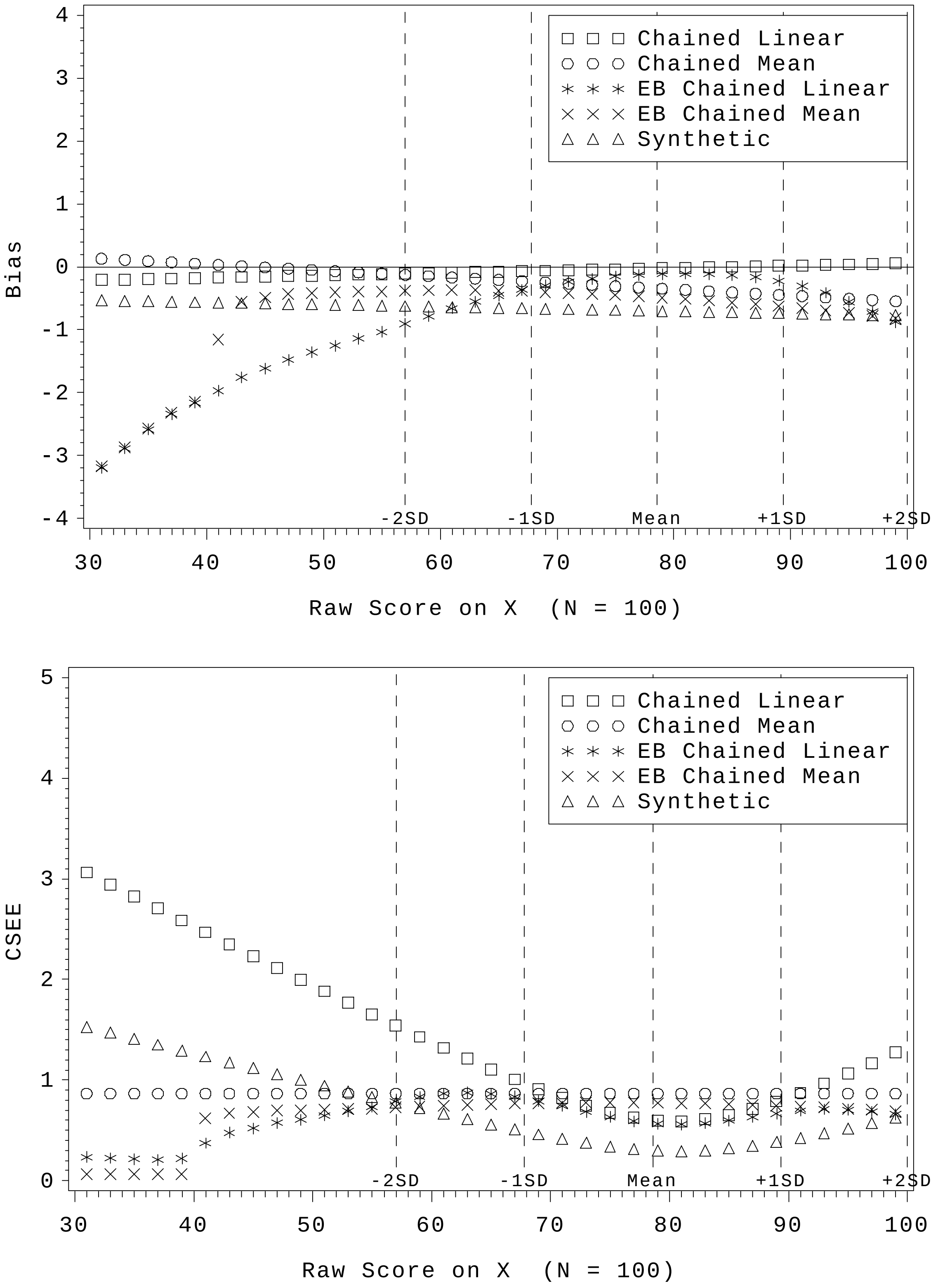


Figure A5. **Conditional bias and conditional standard error of equating (CSEE) at samples of 100 examinees in Study 1.**

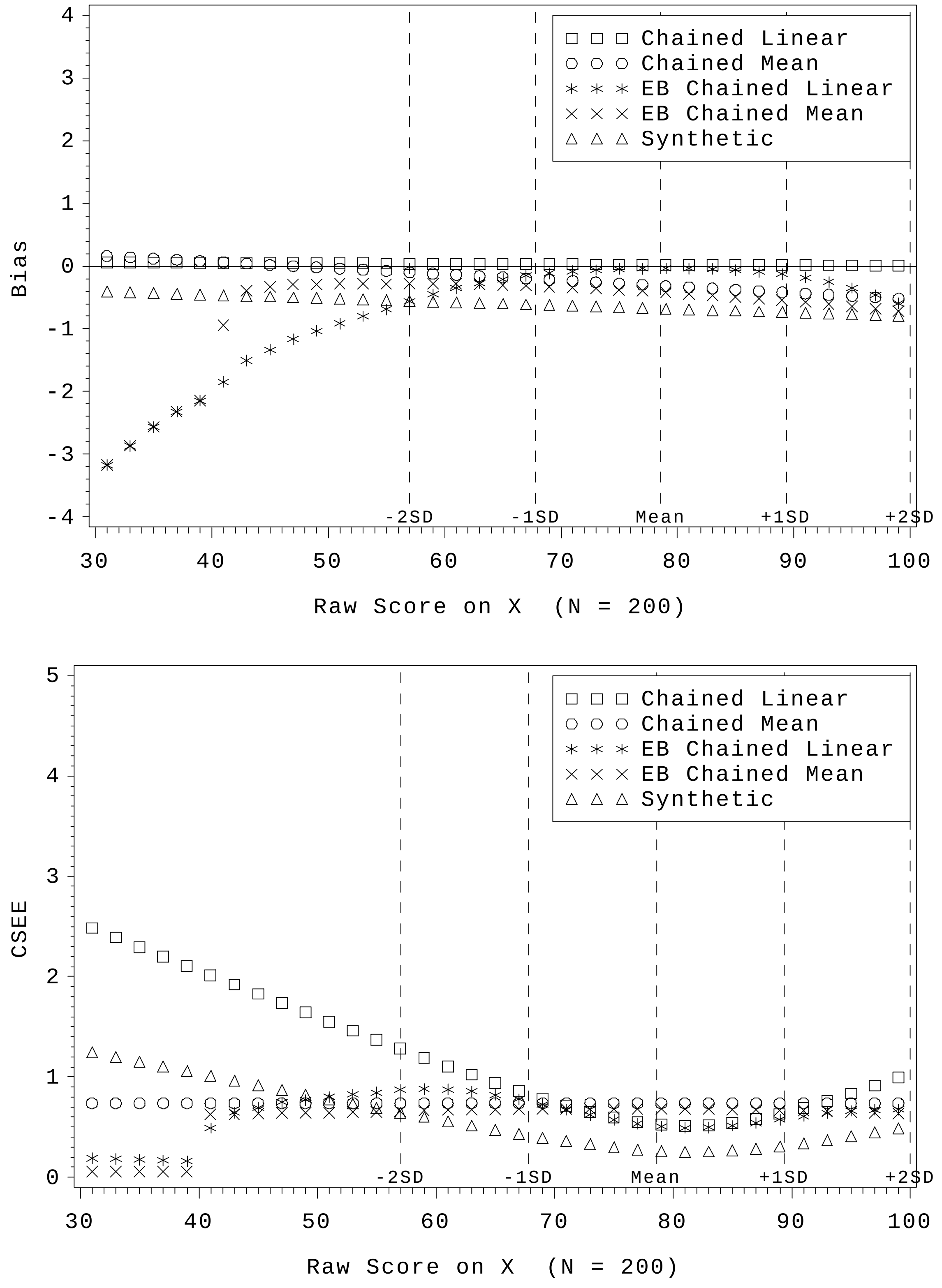
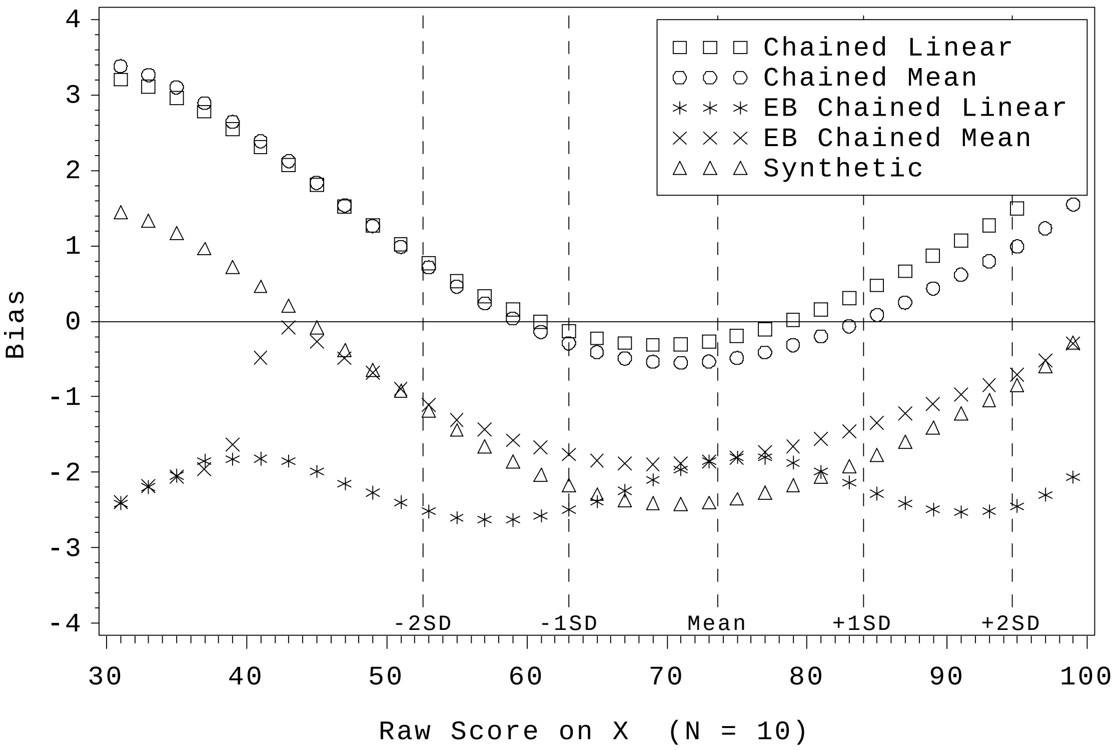


Figure A6. **Conditional bias and conditional standard error of equating (CSEE) at samples of 200 examinees in Study 1.**



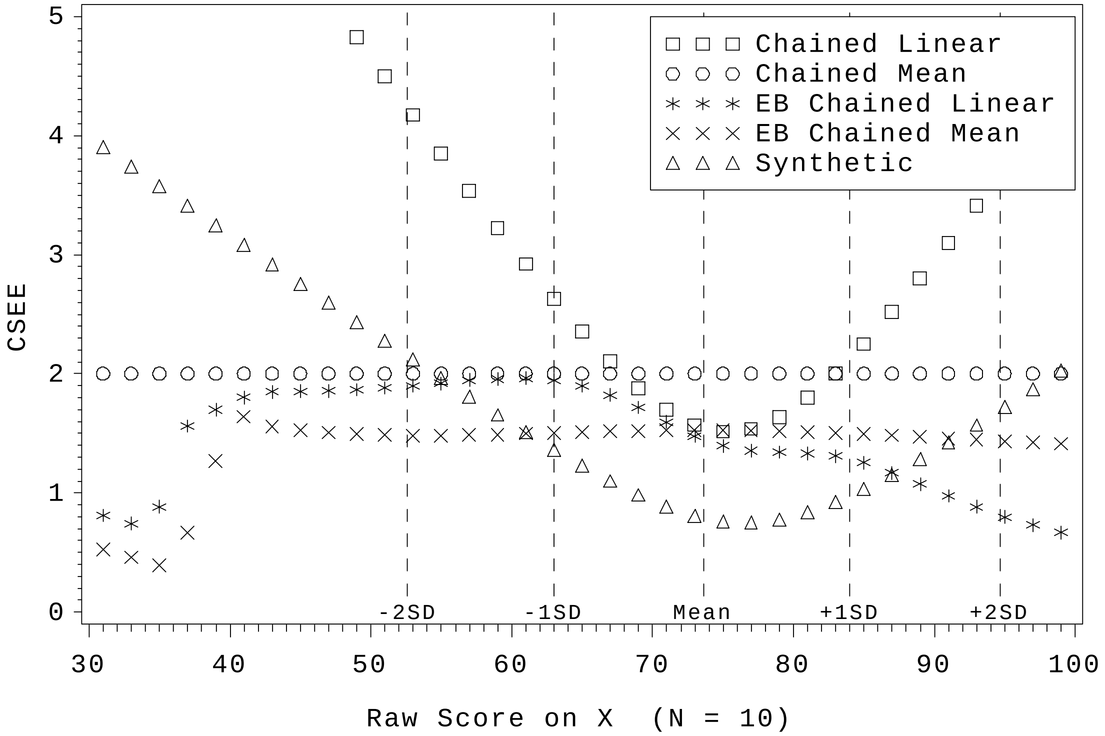


Figure A7. **Conditional bias and conditional standard error of equating (CSEE) at samples of 10 examinees in Study 2.**

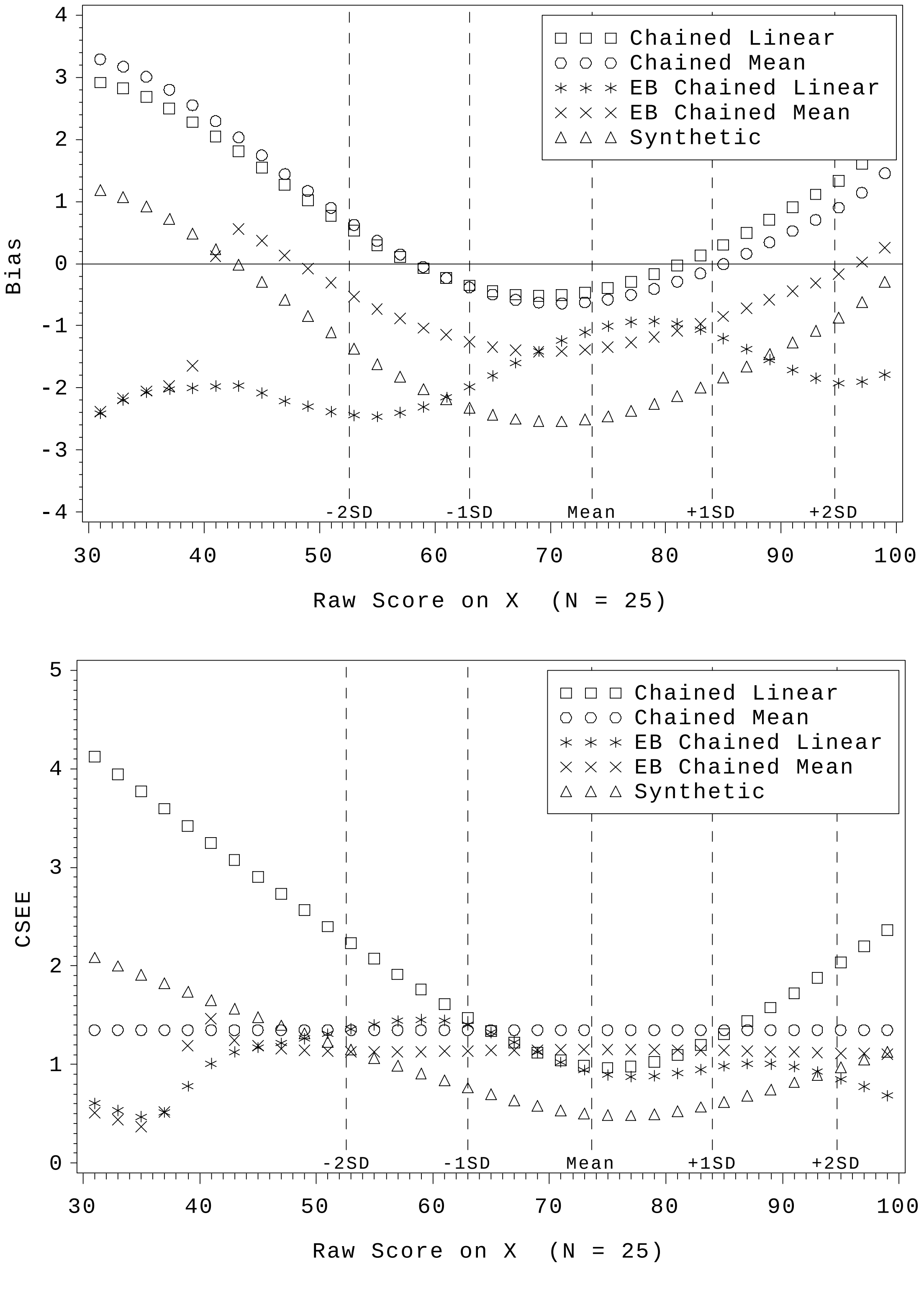
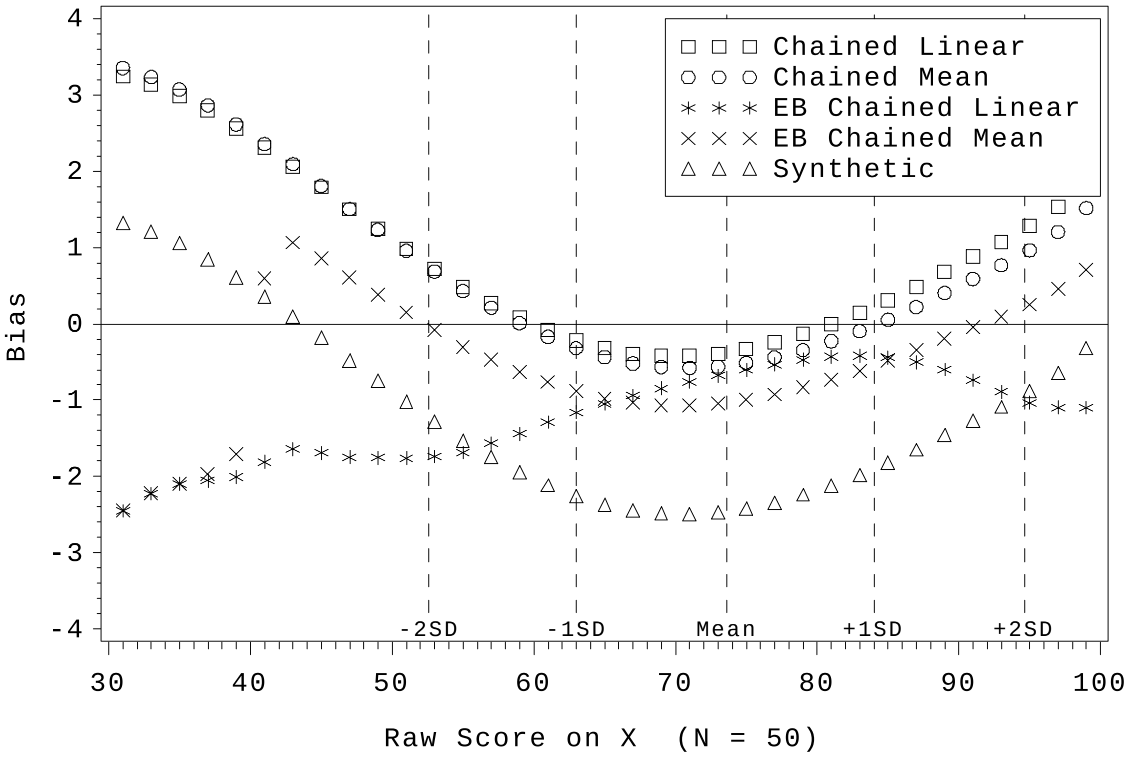


Figure A8. **Conditional bias and conditional standard error of equating (CSEE) at samples of 25 examinees in Study 2.**



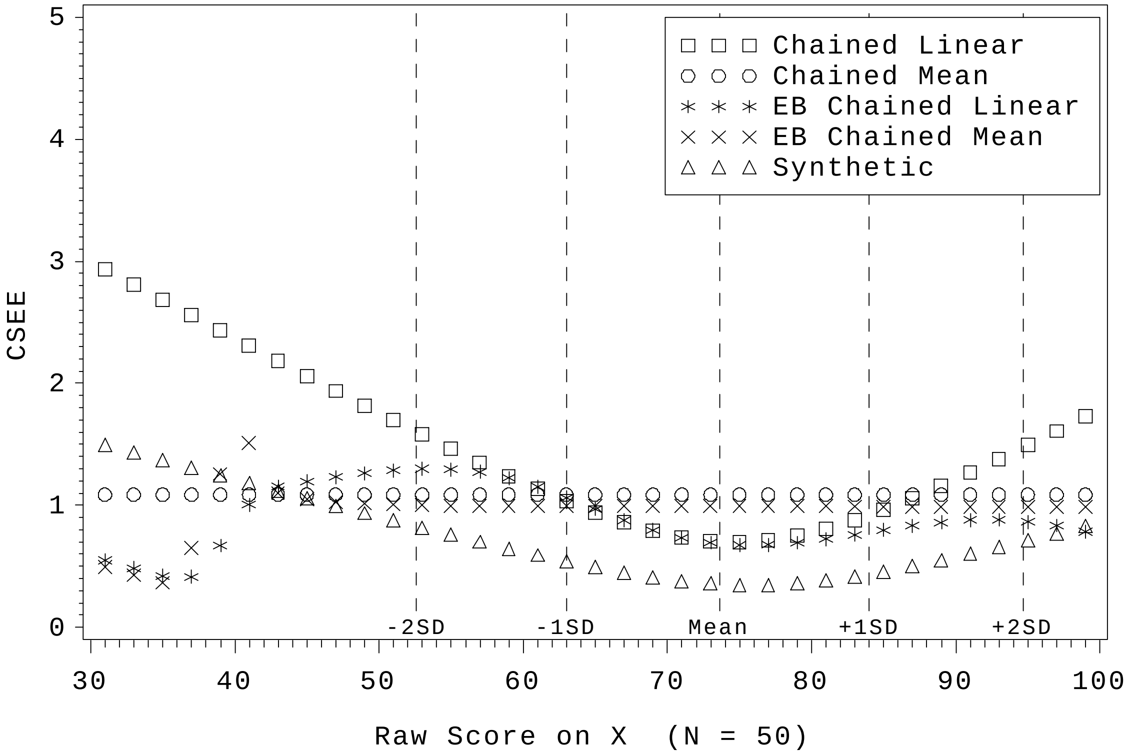
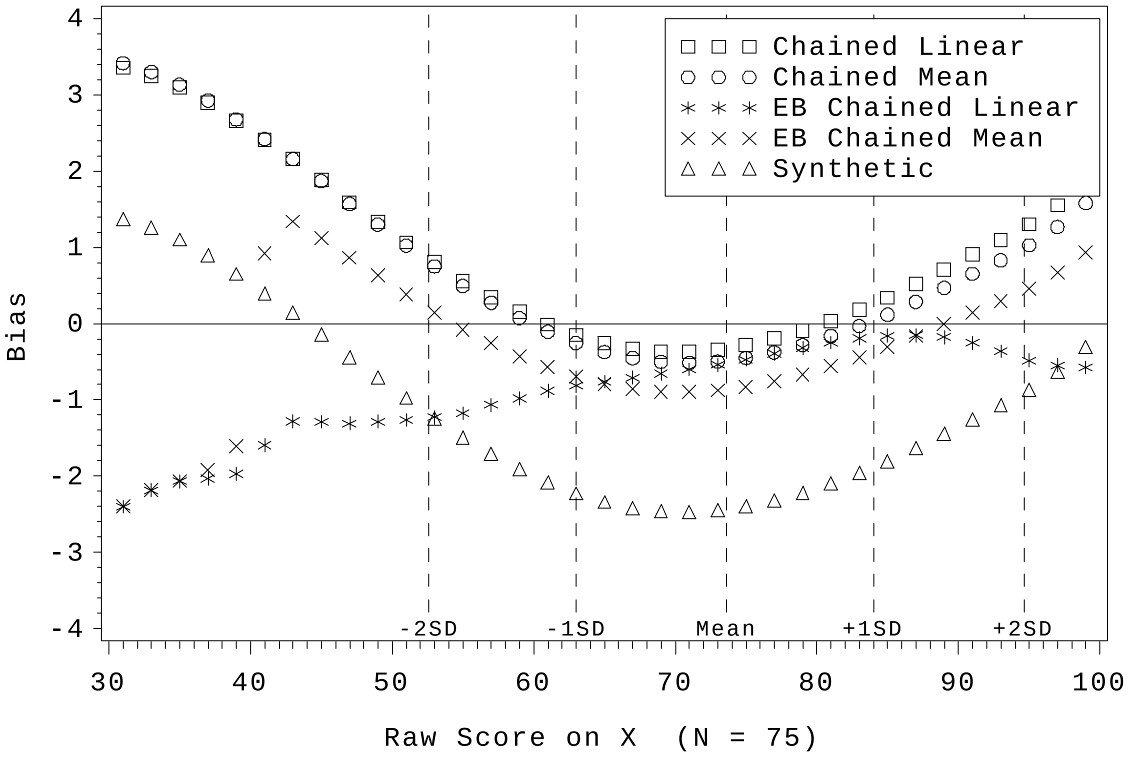


***Figure A9.* Conditional bias and conditional standard error of equating (CSEE) at samples of 50 examinees in Study 2.**



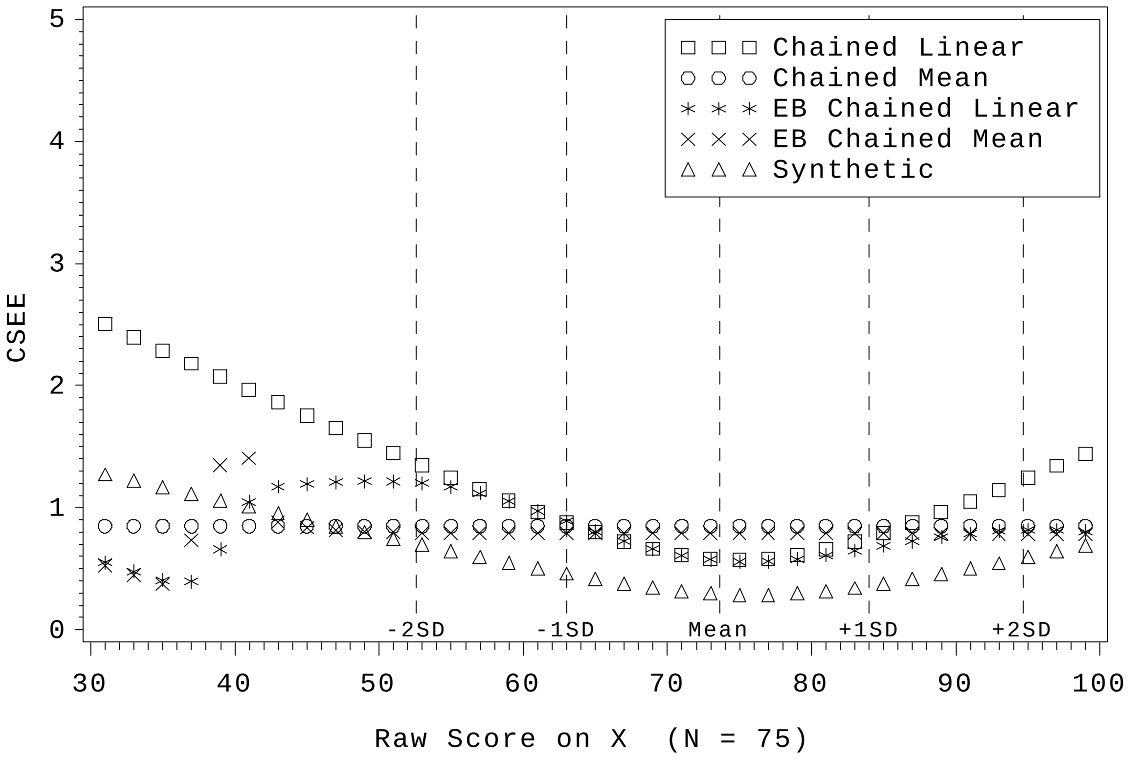


Figure A10. **Conditional bias and conditional standard error of equating (CSEE) at samples of 75 examinees in Study 2.**

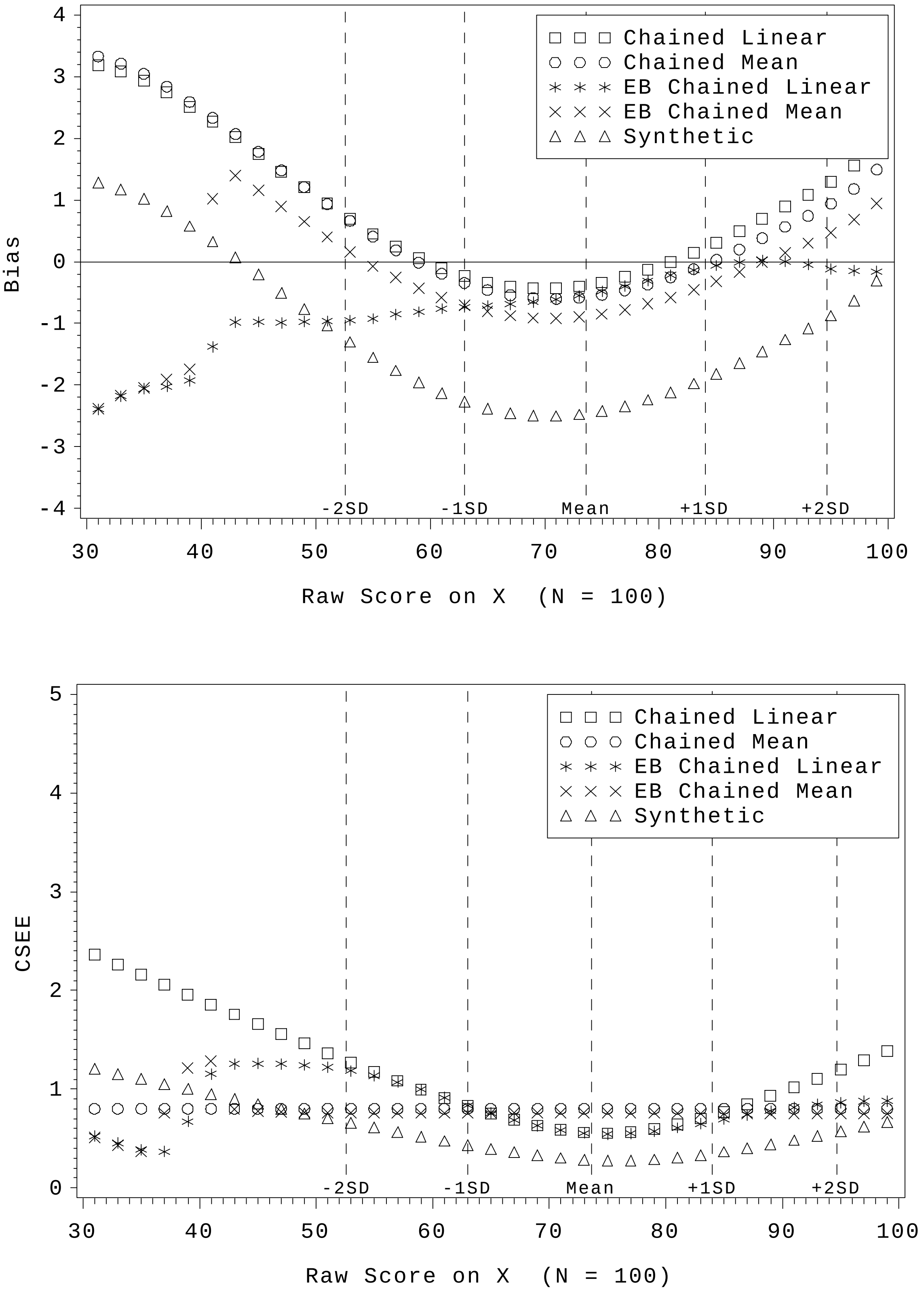


Figure A11. **Conditional bias and conditional standard error of equating (CSEE) at samples of 100 examinees in Study 2.**

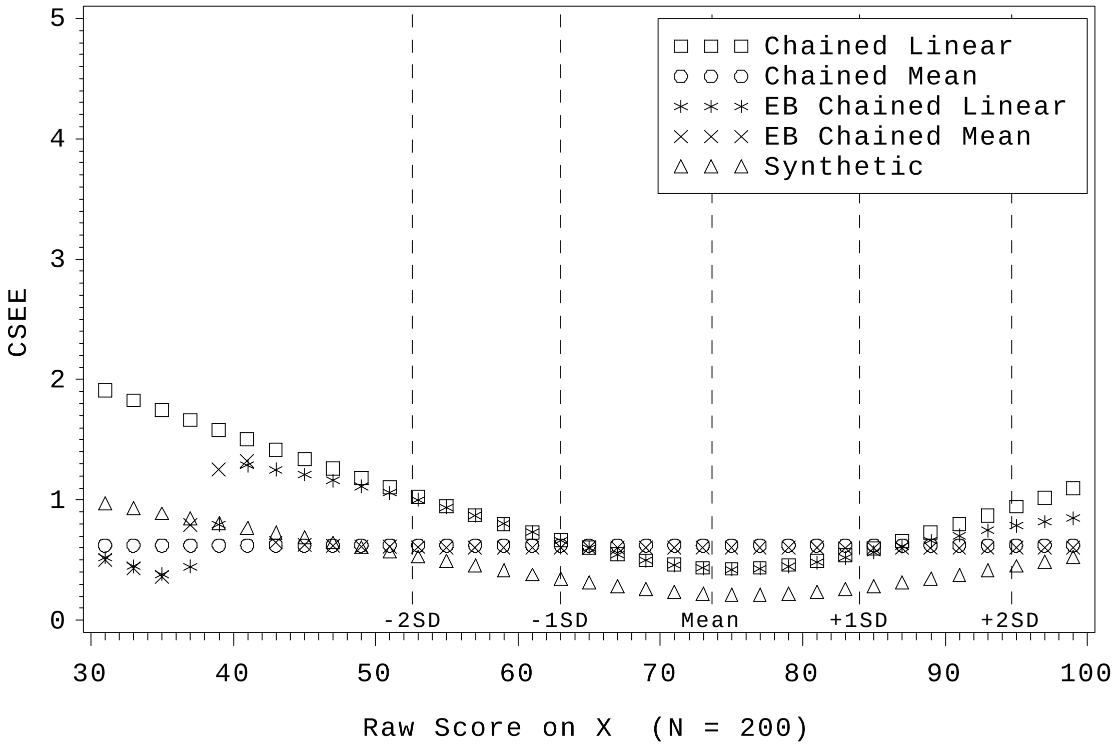


Figure A12. **Conditional bias and conditional standard error of equating (CSEE) at samples of 200 examinees in Study 2.**