Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 19

A Modified Frequency Estimation Equating Method for the Common-Item Non-Equivalent Groups Design

Tianyou Wang
Robert L. Brennan [†]

August 2006

[†] Tianyou Wang is Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: tianyou-wang@uiowa.edu). Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, CASMA, University of Iowa.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

Frequency estimation (also called post-stratification) is an equating method employed under the common-item nonequivalent groups design. A modified frequency estimation method is proposed here, based on altering one of the traditional assumptions in frequency estimation in order to correct for equating bias. A simulation study was carried out to compare equating errors for the modified frequency estimation method, the traditional frequency estimation method, and the chained equipercentile method. The results show that the modified frequency estimation method performs better than the traditional frequency estimation method, and under most circumstances, also performs better than the chained equipercentile method.

1 Introduction

Under the common-item nonequivalent groups design, the new test form (designated as Form X) and the old test form (designated as Form Y) are administered to two different groups of examinees. An anchor test (designated as V) is administered to both of the groups taking the new and the old forms. The data for the anchor test V are used to link the scores on the new test form to the old test form. Under the common-item nonequivalent groups design, there are two different equipercentile equating methods: the frequency estimation (FE) method (also called the post-stratification equating method) and the chained equipercentile (CE) method. For the FE method, the score distributions of Form X and Form Y for a common synthetic population are estimated based on assumptions about the conditional score distributions of X give V and Y given V, and then equipercentile equating is applied to compute the equating relationship. For the CE equating method, Form X scores are first equated to the common-item V scores using the equipercentile equating method. Then, the equated V equivalents of Form X scores are equated to Form Y scores.

Previous studies that compared these two equipercentile methods using real test data (Braun & Holland, 1982; Harris & Kolen, 1990; Livingston, Dorans, & Wright, 1990; Marco, Petersen, & Stewart, 1983) found that these two equating methods generally produced quite different results. These studies also suggest that the CE method may be a better choice when the two groups differ substantially, although the FE method appears to have better theoretical standing. von Davier, Holland and Thayer (2004a) showed that they are both examples of what they termed observed-score equating and they both produce the same equating function when (1) the two populations are very similar or (2) the anchor test is perfectly correlated with both tests. Holland, von Davier, Sinharay, and Han (2006) used real data to test predictions from the assumptions of the CE and FE methods and found that the CE method makes slightly more accurate predictions. Wang, Lee, Brennan and Kolen (2006) used simlulated data to compare the equating errors of these two methods in a more comprehensive manner and found that FE method generally has larger equating bias than the CE method, and the bias is larger when there are larger group differences. The FE method, however, has smaller standard errors of equating (SEE) than the CE method. The comparison of overall equating errors, defined by root mean squared errors (RMSE), depends on the magnitude of group differences. When there are large group differences, the CE method performs better in terms of RMSE; otherwise, the FE performs better.

The purposes of this paper are: (1) to examine the reasonableness of the assumption of the FE method under certain measurement models in order to illuminate the potential cause of equating bias often found with the FE method; (2) to propose a modified frequency estimation method to correct for equating bias; (3) to compare the modified method with the traditional FE method and the CE method using simulation conditions similar to those used in Wang, Lee, Brennan and Kolen (2006).

2 An Analysis of the Basic Assumption of the Frequency Estimation Method under Two Measurement Models

Assume Form X is administered in Population 1, Form Y is administered in Population 2 and the anchor set of items V is administered in both populations. Let X and Y denote the random variable for test scores on Form X and Form Y, respectively. The basic assumption of the FE method is that the conditional distribution of X (or Y) conditional on$V$ remains the same in Populations 1 and 2 (see Kolen & Brennan 2004). Based on this assumption, the joint distribution of X and V, and subsequently the marginal distribution of X in Population 2, are estimated. The marginal distributions of X in both Populations 1 and 2 are then used to form the synthetic population distribution of X. A similar process is followed to obtain the synthetic population distribution for Y. The following discussion shows that this basic assumption of the FE method is violated under both a congeneric model and an IRT model. To simplify the problem, only the conditional mean is analyzed because if the assumption does not hold for the conditional mean, it certainly does not hold for the full conditional distribution.

2.1 Congeneric Model

If X and V follow a congeneric model,

$$X = \lambda_X T + E_X, \tag{1}$$

$$V = \lambda_V T + E_V, \tag{2}$$

where T is the true score, λ_X and λ_V are the congeneric coefficients for X and V, E_X and E_V are errors scores. Under the assumption of homoscedasticity, we can express the conditional mean of X for Population 1 as

$$\mu_1(X|V) = \mu_1(X) + \rho(X,V)\frac{\sigma(X)}{\sigma(V)}[V - \mu_1(V)], \tag{3}$$

where μ_1 designates a mean, σ designates standard deviation, and $\rho(X,V)$ is the correlation between X and V. Here we assume that $\rho(X,V)$, $\sigma(X)$, and $\sigma(V)$ are the same for both populations, so subscripts are omitted for these terms. Assume that there is a difference δ in the true score of the anchor set between Population 1 and Population 2; that is

$$\mu_2(V) - \mu_1(V) = \lambda_V T_2 - \lambda_V T_1 = \delta. \tag{4}$$

Then we have

$$\mu_2(X) - \mu_1(X) = \frac{\lambda_X}{\lambda_V}\delta. \tag{5}$$

Now, using Equations 4 and 5 in the definition of $\mu_2(X|V)$ gives:

$$\mu_2(X|V) \quad = \quad \mu_2(X) + \rho(X,V)\frac{\sigma(X)}{\sigma(V)}[V - \mu_2(V)]$$

$$= \quad \mu_1(X) + \frac{\lambda_X}{\lambda_V}\delta + \rho(X,V)\frac{\sigma(X)}{\sigma(V)}[V - \mu_1(V) - \delta]. \tag{6}$$

Subtracting Equation 3 from Equation 6, we have

$$\mu_2(X|V) - \mu_1(X|V) = \left[\frac{\lambda_X}{\lambda_V} - \rho(X,V)\frac{\sigma(X)}{\sigma(V)}\right]\delta. \tag{7}$$

Further,

$$\begin{aligned} \rho(X,V)\frac{\sigma(X)}{\sigma(V)} &= \frac{\lambda_X\lambda_V\sigma^2(T)}{\sigma(X)\sigma(V)}\;\frac{\sigma(X)}{\sigma(V)} \\ &= \frac{\lambda_X}{\lambda_V}\;\frac{\lambda_V^2\sigma^2(T)}{\sigma^2(V)} \\ &= \frac{\lambda_X}{\lambda_V}\,\rho(V,V'), \end{aligned} \tag{8}$$

where $\rho(V,V')$ is the reliability of V.

Thus, Equation 7 can be expressed as

$$\mu_2(X|V) - \mu_1(X|V) = \frac{\lambda_X}{\lambda_V}\,\delta\,[1 - \rho(V,V')]. \tag{9}$$

Clearly, the conditional mean will not remain constant from Population 1 to Population 2 unless the anchor set has perfect reliability. This means that under the realistic situation where anchor set does not have perfect reliability, the basic assumption of the FE method is violated.

2.2 IRT Model

It is difficult to show analytically that the assumptions of the FE method generally do not hold under the IRT model. In order to show that a certain assumption does not always hold, however, it suffices to show that it does not hold for some special case. The first 60-item mathematics test form used in Wang, Lee, Brennan, and Kolen (2006) is used to examine whether the conditional means change from one population to another. For Population 1, a normal distribution N(0,1) is used. For Population 2, a N(.25,1) is used. A procedure similar to the one described in Wang, Kolen and Harris (2000) is used to compute the bivariate distribution of X and V under each population. The conditional distributions and conditional means are subsequently computed. Table 1 contains the conditional means for Form X for both populations and the differences. The non-zero differences show that the basic assumption of the FE method is violated in this case.

2.3 Summary

Based on the above analyses for these two different measurement models, we may conjecture that the basic assumption of the FE method is violated under any

Table 1: The Conditional Means for Two Different Populations

Score	Population 1	Population 2	Differences
0	8.168902	8.354938	0.186035
1	9.631878	9.852101	0.220223
2	11.197180	11.452809	0.255629
3	12.886189	13.176914	0.290724
4	14.722144	15.045357	0.323213
5	16.725791	17.075785	0.349994
6	18.908108	19.275773	0.367665
7	21.261834	21.635631	0.373798
8	23.757689	24.126263	0.368574
9	26.351241	26.706517	0.355277
10	28.998046	29.336791	0.338745
11	31.666613	31.989692	0.323079
12	34.342250	34.652796	0.310546
13	37.024257	37.326160	0.301903
14	39.721582	40.018668	0.297086
15	42.448439	42.743944	0.295505
16	45.218732	45.514702	0.295970
17	48.039403	48.336001	0.296598
18	50.905236	51.199992	0.294756
19	53.791998	54.077610	0.285613
20	56.615121	56.870863	0.255742

reasonable measurement model. That is, we may hypothesize that if simulations or analyses were performed under measurement models different from those discussed in this section, similar equating bias would be observed. The following section proposes a modification to the FE method with the goal of reducing bias.

3 A Modified Frequency Estimation Method

The modified frequency estimation (MFE) method attempts to reduce equating bias by modifying the basic assumption of the FE method. Instead of assuming that the distribution of X conditional on V remains invariant across populations, as expressed below

$$f_1(X|V) = f_2(X|V), \tag{10}$$

the MFE method assumes that the distribution of X conditional on the *true score* of V remains invariant across populations. That is, the MFE method assumes that

$$f_1(X|T_V) = f_2(X|T_V). \tag{11}$$

The reason for making this modification in the basic assumption is that we have shown that when we condition on anchor test observed scores, the conditional distribution of X will *not* remain invariant. On the other hand, there are certainly situations in which conditioning on the anchor test *true score* guarantees that the conditional distribution of X will remain invariant. For example, under a congeneric model, conditioning on the true score of the anchor test is the same as conditioning on the true score of X. Also, the distribution of X conditional on the true score for X is identical to the distribution of error scores for X conditional on the true score for X. It follows that, in this case, we are assured that the conditional mean will remain invariant across populations. It is a rather weak assumption to assume that the error scores are distributed the same across two populations. This suggests that the modified assumption is a rather weak assumption and is expected to be met at least approximately under most testing situations.

This assumption is not directly useful, however, because we do not immediately have the distribution of observed (or error) scores conditional on true score. However, we can use a certain relationship between true scores and observed scores to replace the two occurrences of T_V in Equation 11 with observed scores for V, so that we have

$$f_1(X|V_1) = f_2(X|V_2), \tag{12}$$

for V_1 and V_2. The observed data provides $f_1(X|V_1)$ directly. To obtain $f_2(X|V_2)$, for every V_2 we need to find the corresponding V_1. To do so, we adopt Brennan and Lee's (2006) approach to estimating true score from observed score.[1] Their approach applied to the present situation results in the following expressions:

$$T_{V_1} = \mu_1(V) + \sqrt{\rho_1(V,V')}\,[V_1 - \mu_1(V)] \tag{13}$$

and

$$T_{V_2} = \mu_2(V) + \sqrt{\rho_2(V,V')}\,[V_2 - \mu_2(V)], \tag{14}$$

where $\rho_1(V,V')$, and $\rho_2(V,V')$ are the reliabilities of V in the two populations. By setting $T_{V_1} = T_{V_2}$, for every V_2 we can compute the corresponding V_1, namely,

$$V_1 = \frac{\sqrt{\rho_2(V,V')}}{\sqrt{\rho_1(V,V')}}V_2 + \frac{1-\sqrt{\rho_2(V,V')}}{\sqrt{\rho_1(V,V')}}\,\mu_2(V) - \frac{1-\sqrt{\rho_1(V,V')}}{\sqrt{\rho_1(V,V')}}\,\mu_1(V). \tag{15}$$

Then, Equation 12 can be used to find $f_2(X|V_2)$. Note, however, that for a given integer score V_2, the corresponding V_1 will likely be a non-integer. If we have a continuous bivariate distribution of X and V_1, then the non-integer

[1] The basic idea is to find a linear transformation of observed scores to estimated true scores such that the estimates have a variance equal to true score variance. An alternative procedure would be to use Kelley's (1947) regressed score estimates, but various analyses have demonstrated to the authors that this alternative does not work as well as the Brennan and Lee (2006) procedure for the situation considered in this paper.

value will not be an issue. Usually, however, we have only a discrete bivariate distribution, and can compute $f_1(X|V_1)$ only for integer V_1. In this case, we can use linear interpolation between two adjacent integer scores around V_1 to compute $f_1(X|V_1)$ for non-integer V_1. The conditional distribution obtained from linear interpolation should be normalized so that the sum of the conditional probabilities over all the possible discrete scores equals one. After we obtain $f_2(X|V_2)$ for all integer V_2 scores, we can compute the marginal distribution of X in Population 2 as follows:

$$f_2(X) = \sum_{V_2} f_2(X|V_2) f_2(V_2). \tag{16}$$

The marginal distribution of X in the synthetic population is:

$$f_S(X) = w_1 f_1(X) + w_2 f_2(X), \tag{17}$$

where w_1 and w_2 are the synthetic population weights for Populations 1 and 2, respectively. The synthetic population marginal distribution of Y, $f_S(Y)$, can be found in a similar manner. The remaining equating steps are the same as for the traditional FE method (see Kolen & Brennan, 2004).

4 An Illustration with Real Test Data

A real test data set from Brennan and Kolen (2004) is used here to illustrate the MFE method. In this data set, both Form X and Form Y have 36 items, with 12 common items. The sample size for Form X is 1655 and for From Y is 1638. The CE, FE, and MFE are employed. In order to display the differences among these methods, the equating functions minus the identity function are plotted, as is done frequently in Kolen and Brennan (2004). Figure 1 shows that throughout most of the score scale, the MFE equivalents lie between the FE and CE equivalents, which means the modification of the FE method proposed here decreases the difference between the frequency estimation method (in the general sense) and the chained-equipercentile method.

5 Simulation Study

A simulation study was conducted to compare the modified FE method with the traditional FE method and the CE method. The simulation procedure using the IRT model described in Wang, Lee, Brennan, and Kolen (2006) was used in this study. Two pairs of test forms from that study are used here: two 60-item test forms with a 20-item anchor set and two 120-item test forms with a 24-item anchor set. The item parameter estimates for these two pairs of test forms are taken as true item parameters and are used to generate item responses given persons' θ parameters. The descriptive statistics for the item parameters for these two pairs of test forms are in Tables 2 and 3. The mean difference between Form X and Form Y in item difficulty parameters for the 60-item forms

Table 2: Descriptive Statistics of Item Parameters For the 60-Item Test Forms with 20-Item Anchor Test

	Parameter	n	Mean	SD	Skewness	Kurtosis
Common Items						
	a	20	1.041889	0.289427	0.369885	1.789235
	b	20	0.240455	0.919670	-0.376598	2.704618
	c	20	0.145658	0.043153	0.033088	3.256947
Form X						
	a	60	1.035269	0.258278	0.379614	2.191785
	b	60	0.324333	0.931717	-0.256400	2.390305
	c	60	0.148527	0.049444	-0.051281	2.125785
Form Y						
	a	60	0.992131	0.302323	0.257331	2.065286
	b	60	0.233642	0.942919	-0.295781	2.417456
	c	60	0.154173	0.043062	0.101103	2.683931

is about 0.09, which represents a relatively small difference in difficulty. The mean difference in item difficulty parameters for the 120-item forms is about 0.17, which represents a moderate difference in difficulty.

The “true” equating function is defined and computed using the same procedure as in Wang, Lee, Brennan, and Kolen (2006); that is, the IRT observed score equating function, given a standard normal θ distribution for both test forms, is used as the “true” equating function.

Random samples of θ values are drawn from two different θ distributions—one associated with Form X and one with Form Y—to simulate the common-item nonequivalent groups design. The θ distribution for Form X is always set as $N(0, 1)$. Six variations of θ distribution for Form Y are used: $N(0.05, 1)$, $N(0.1, 1)$, $N(0.25, 1)$, $N(0.25, 1.2)$, $N(-0.1, 1)$, and $N(0.1, 0.8)$. These six variations represent different degrees and directions of group differences in mean and variance. The sample size for the 60-item pair is 2000 and the sample size for the 120-item pair is 4000.

After a pair of random samples are drawn, the simulees’ responses to test items are generated using the 3PL IRT model, and their test scores are computed. Equating functions are computed using the different equating methods considered here (FE, CE,and MFE) and are compared to the “true” equating function to compute the equating errors. This process is replicated 500 times. The same evaluation criteria used in Wang, Lee, Brennan, and Kolen (2006) are used here: conditional equating bias, standard error of equating (SEE), and root mean squared error (RMSE) conditional on each score point; and aggregate equating bias, SEE, and RMSE. The aggregate error indexes are weighted averages of the conditional error indexes weighted by the Form X population score distribution relative frequencies. For bias, the absolute value is used in

Table 3: Descriptive Statistics of Item Parameters For the 120-Item Test Forms with 24-Item Anchor Test

	Parameter	n	Mean	SD	Skewness	Kurtosis
Common-Items						
	a	24	0.918253	0.329397	0.653700	2.758182
	b	24	0.131234	1.075756	-0.311878	2.345994
	c	24	0.147795	0.041002	-0.124009	2.508089
Form X						
	a	120	0.997865	0.288740	0.184943	2.414850
	b	120	0.253162	0.958518	-0.286780	2.408884
	c	120	0.151625	0.045525	-0.027105	2.312946
Form Y						
	a	120	0.966850	0.318308	0.208014	2.286994
	b	120	0.088879	0.941646	-0.538595	2.859228
	c	120	0.141880	0.042935	-0.142998	2.792031

computing the weighted average.

This simulation study is not intended to study the effect of a large number of factors in a systematic manner. Rather, it is designed to provide variation in a sufficient number of conditions to provide a reasonable, initial comparison of the characteristics of the MFE method relative to the FE and CE methods.

6 Results

The conditional equating bias and SEE of the four equating methods are plotted in Figures 2 through 7. The aggregate error indexes are in Tables 4 and 5. They are discussed separately below.

6.1 Conditional Equating Errors

Figures 2 through 4 are for the 60-item forms. Figure 2 shows that when there is a small group mean difference ($\mu_Y = .05$), the bias of the FE method is only slightly larger than for the CE method, and the bias reduction for the MFE method over the FE method is small, too. The RMSE for the MFE method is about the same as for the FE method. When there is a moderate group mean difference ($\mu_Y = .1$), the bias reduction for the MFE method is more apparent. Because the MFE method has about the same SEE as the FE method, the net effect is that the MFE method has a slightly lower RMSE than the FE method. For both the small and moderate group mean difference cases, the RMSEs for the FE and MFE methods are smaller than for the CE method.

The left side of Figure 3 shows that when there is a large group mean difference ($\mu_Y = .25$), the MFE method has considerably lower bias than the FE method. The reduction of bias is most salient in the middle and upper parts of the score scale. RMSE for the MFE method is not only lower than for the FE method, but also lower than that for the CE method. The right side of Figure 3 shows that when there is a combined group mean and variance difference ($\mu_Y = .25$, $\sigma_Y = 1.2$), the MFE method greatly reduces bias relative to the FE method, especially at the upper part of the score scale. The MFE method has lower RMSE than the FE method all along the score scale. The MFE method has lower RMSE than the CE method in the middle and upper parts of the score scale. At lower parts of the score scale, the MFE method sometimes has a slightly higher RMSE than the CE method.

The left side of Figure 4 shows that when there is a moderate group mean difference in the opposite direction (i.e., $\mu_Y = -.1$), the bias for all the methods is negative, but the shape of the bias curves remains unchanged. The MFE method has slightly lower bias and RMSE than the FE method. The MFE method has lower RMSE than the CE method. The right side of Figure 4 shows that when there is a variance difference in the opposite direction (i.e., $\mu_Y = .1$, $\sigma_Y = .8$), the MFE method greatly reduces bias relative to the FE method all along the score scale. Bias for the MFE method is slightly larger than for the CE method. RMSE for the MFE method is much lower than for the FE method. RMSE for the MFE method is slightly lower than for the CE method in the middle and upper parts of the score scale, but is somewhat higher than for the CE method in the lower part of the score scale.

The same pattern of bias reduction for the MFE method over the FE method is observed for the 120-item forms in Figures 5 through 7. The amount of bias reduction is even more salient than for the 60-item forms even when there is only a small group mean difference. This may be due to the larger form difference in terms of difficulty, or it may be due to the longer test length.

Overall, the conditional error indexes show that the MFE method performs quite well in reducing bias relative to the FE method. The SEEs for the the MFE and FE methods are about the same. In all cases we examined, the MFE method has lower RMSE than the FE method. In most cases, the FME method has lower RMSE than the CE method. The only case in which the CE method has lower RMSE than the MFE method is when Population 1 has a larger variance than Population 2 ($\mu_Y = .1$, $\sigma_Y = .8$).

6.2 Overall Equating Errors

The results for the aggregate equating errors, as shown in Tables 4 and 5, basically confirm the general comparative pattern in the conditional error indexes. In terms of bias and RMSE, the MFE method performs better than the FE method in all cases. In terms of RMSE, the MFE method performs better than the CE method in all cases except when Population 1 has a larger variance than Population 2 ($\mu_Y = .1$, $\sigma_Y = .8$). In that case, the CE method has only slightly smaller RMSE than the MFE method.

Table 4: Aggregate Equating Errors For the 60-Item Test Forms with 20-Item Anchor Test

Index	CE	FE	MFE
$(\mu_Y = 0.05, \sigma_Y = 1.0)$			
Abs. Bias	0.032253	0.060970	0.030667
SEE	0.364353	0.313860	0.316044
RMSE	0.366256	0.320611	0.318104
$(\mu_Y = 0.1, \sigma_Y = 1.0)$			
Abs. Bias	0.065229	0.125692	0.057273
SEE	0.363385	0.314139	0.315879
RMSE	0.370152	0.340583	0.323260
$(\mu_Y = 0.25, \sigma_Y = 1.0)$			
Abs. Bias	0.166583	0.320030	0.134288
SEE	0.366330	0.315625	0.316458
RMSE	0.405036	0.454302	0.355192
$(\mu_Y = 0.25, \sigma_Y = 1.2)$			
Abs. Bias	0.178853	0.338050	0.132415
SEE	0.368165	0.320865	0.321725
RMSE	0.418784	0.485298	0.356545
$(\mu_Y = -0.1, \sigma_Y = 1.0)$			
Abs. Bias	0.072296	0.135481	0.052803
SEE	0.363314	0.313343	0.315094
RMSE	0.371148	0.343085	0.322647
$(\mu_Y = 0.1, \sigma_Y = 0.8)$			
Abs. Bias	0.268614	0.486309	0.343942
SEE	0.365752	0.312809	0.317445
RMSE	0.473522	0.602552	0.495001

7 Conclusions and Discussion

Based on the above results, the MFE method performs better than the FE method. Further, the only condition under which the MFE method does not have a clear cut advantage over the CE method is when the new group has larger variance than the old group. Even in that case, however, the MFE probably should be favored unless the focus of attention is on the lower part of the score scale. In short, the results in this paper generally support use of the MFE method over the FE or CE methods.

Some caveats are warranted. First, since the simulation study uses an IRT model for generating response data, the applicability of the results of the simulation study to real world testing situations rests upon the goodness of fit of the IRT model to real test data. Second, the definition of the "true" equating function has a direct bearing on the quantification of bias. Wang, Lee, Brennan,

Table 5: Aggregate Equating Errors For the 120-Item Test Forms with 24-Item Anchor Test

Index	CE	FE	MFE
$(\mu_Y = 0.05, \sigma_Y = 1.0)$			
Abs. Bias	0.100859	0.175520	0.082090
SEE	0.551226	0.470760	0.475202
RMSE	0.562226	0.505595	0.484707
$(\mu_Y = 0.1, \sigma_Y = 1.0)$			
Abs. Bias	0.193290	0.344705	0.145654
SEE	0.552046	0.471449	0.475717
RMSE	0.588280	0.589591	0.505158
$(\mu_Y = 0.25, \sigma_Y = 1.0)$			
Abs. Bias	0.463164	0.844723	0.345424
SEE	0.553170	0.473355	0.476823
RMSE	0.729987	0.975770	0.614089
$(\mu_Y = 0.25, \sigma_Y = 1.2)$			
Abs. Bias	0.523872	0.938047	0.396147
SEE	0.561484	0.486681	0.488017
RMSE	0.796552	1.092306	0.649226
$(\mu_Y = -0.1, \sigma_Y = 1.0)$			
Abs. Bias	0.173229	0.328322	0.114796
SEE	0.556951	0.476544	0.482267
RMSE	0.584579	0.582332	0.502294
$(\mu_Y = 0.1, \sigma_Y = 0.8)$			
Abs. Bias	0.683819	1.182562	0.806359
SEE	0.550686	0.462021	0.474678
RMSE	0.921655	1.308798	0.981616

and Kolen (2006) provide a justification for the definition of the "true" equating function used here. Third, the regression function used in the the MFE method is largely based on the classical test theory model. The fact that an IRT-based simulation provides supportive evidence for the MFE method can be viewed as a cross validation for the method. Still, it would be desirable that subsequent simulations be conducted to evaluate the MFE method against the FE and CE methods in some classical test theory framework.

This study does not use any smoothing technique in the equating procedures. It would be interesting to compare the performance of the MFE method against the FE and CE methods when smoothing is used. The bivariate log-linear smoothing method could be used in such a study. In order to solve the non-integer issue in the MFE method (see section 3), it would be advantageous to adopt the smoothing/continuization approach proposed by Wang (2005) to obtain a continuous bivariate distribution.

This paper focuses on the conventional percentile rank-based equipercentile method. The modification proposed here to the traditional frequency estimation method can also be applied to the kernel equating method (von Davier, Holland, & Thayer, 2004b). It seems reasonable to speculate that doing so would have the same advantages for the kernel method as for the percentile rank-based equipercentile method.

Recently, von Davier, Fournier-Zajac, and Holland (2006) proposed an equipercentile version of the Levine linear observed-score equating method, which is based on assumptions about true scores. Since the MFE method involves assumptions about distributions conditional on true score, the MFE method seems conceptually close to the Levine method. For this reason, the method proposed by von Davier, Fournier-Zajac, and Holland may be similar at least conceptually to the MFE method. A future study could be carried out using both real test data and simulated data to compare these two methods.

References

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

Brennan, R. L., & Lee, W. (2006). *Correcting for bias in single-administration decision consistency indexes.* (CASMA Research Report No. 18). Iowa City, Iowa: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. Retrieved from http://www.education.uiowa.edu/casma

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods for observed-score equating: their relationship to population invariance. *Journal of Educational Measurement, 41*, 15–32. von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating.* New York: Springer-Verlag.

von Davier A. A., Fournier-Zajac, S., & Holland, P. W. (2006). *An equipercentile version of the Levine Linear Observed-score equating function using the methods of kernel equating.* Paper presented at the annual meeting of National Council of Measurement in Education, April, San Francisco.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26,* 3–24.

Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement, 50*, 61–71.

Holland, P. W., von Davier, A. A., Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and post-stratification equating methods for the NEAT design.* Paper presented at the annual meeting of

National Council of Measurement in Education, April, San Francisco.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

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

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combinations of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73–95.

Marco, G. L, Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp. 147–176). New York: Academic Press.

Wang, T. (2005). *An alternative continuization method to the kernel method in von Davier, Holland and Thayer's (2004) test equating framework.* (CASMA Research Report No. 11). Iowa City, Iowa: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. Retrieved from http://www.education.uiowa.edu/casma

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessment using polytomous IRT. *Journal of Educational Measurement, 37*, 141–162.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2006). *A Comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design.* (CASMA Research Report No. 17). Iowa City, Iowa: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. Retrieved from http://www.education.uiowa.edu/casma

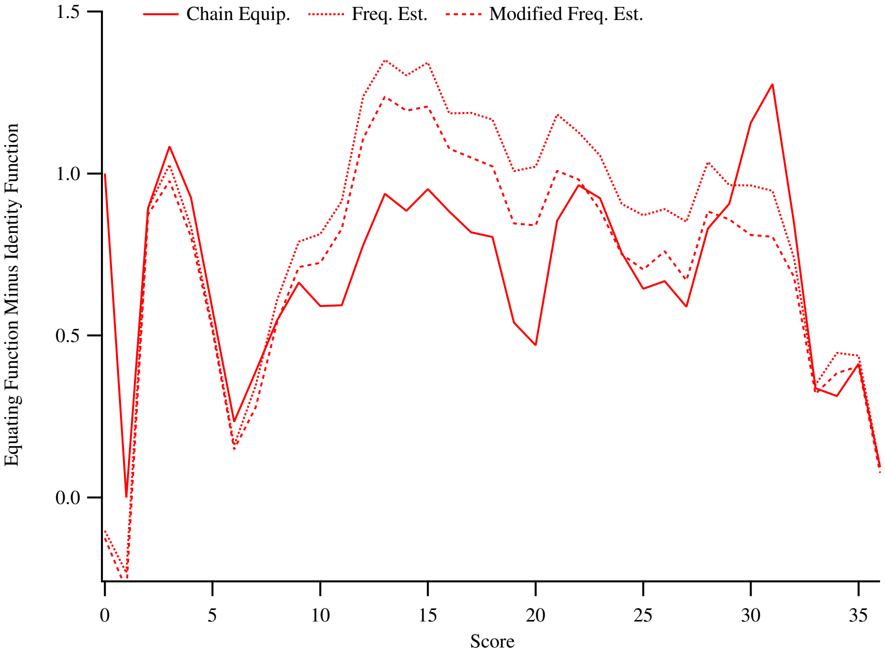

Figure 1: Comparison of CE, FE, and MFE for a 36-Item Real Test Data

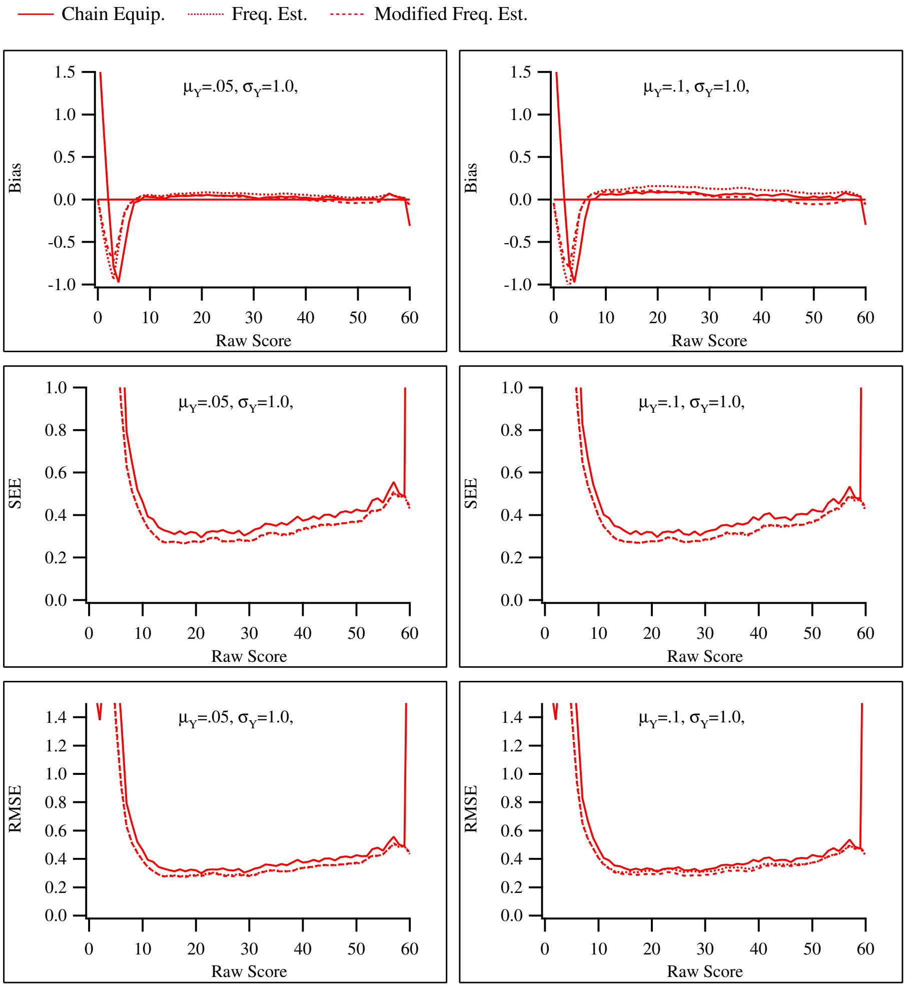

Figure 2: Bias, SEE and RMSE for the 60-Item Test Form Pair with 20-Item Anchor and Equal Group Variances

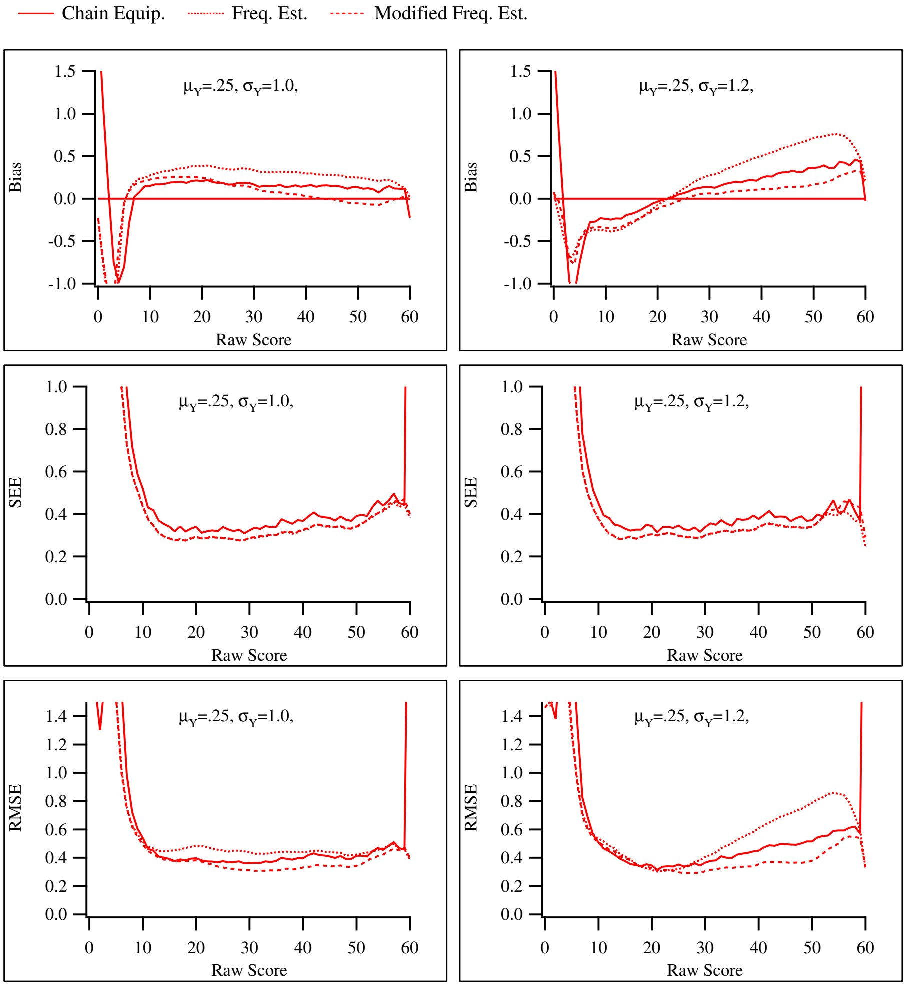

Figure 3: Bias, SEE and RMSE for the 60-Item Test Form Pair with 20-Item Anchor and Unequal Group Variances

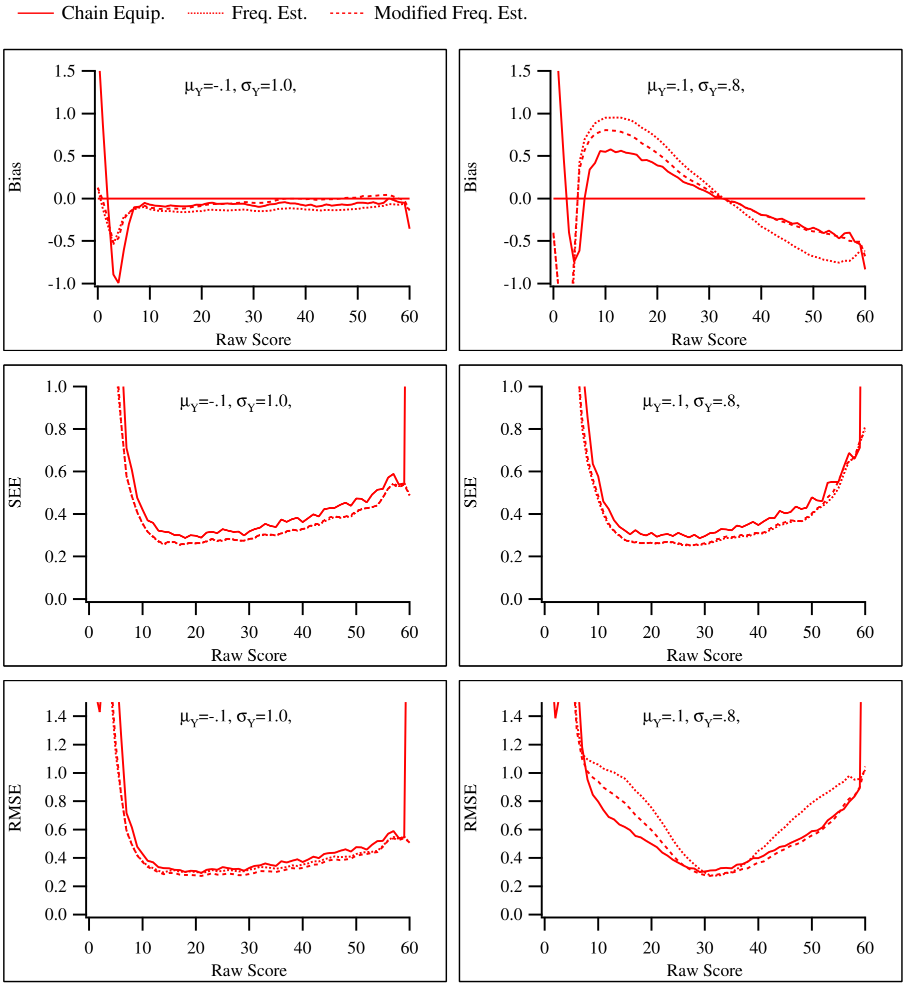

Figure 4: Bias, SEE and RMSE for the 60-Item Test Form Pair with 20-Item Anchor and Group Differences in Opposite Direction

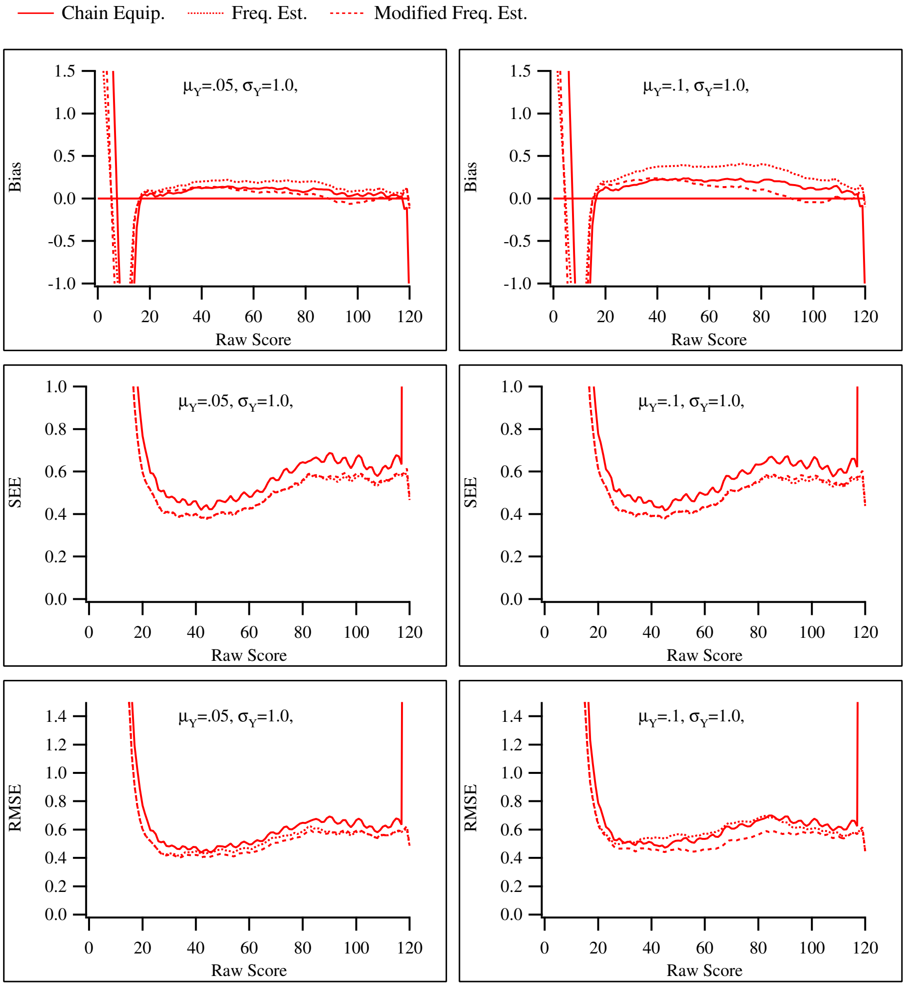

Figure 5: Bias, SEE and RMSE for the 120-Item Test Form Pair with 24-Item Anchor and Equal Group Variances

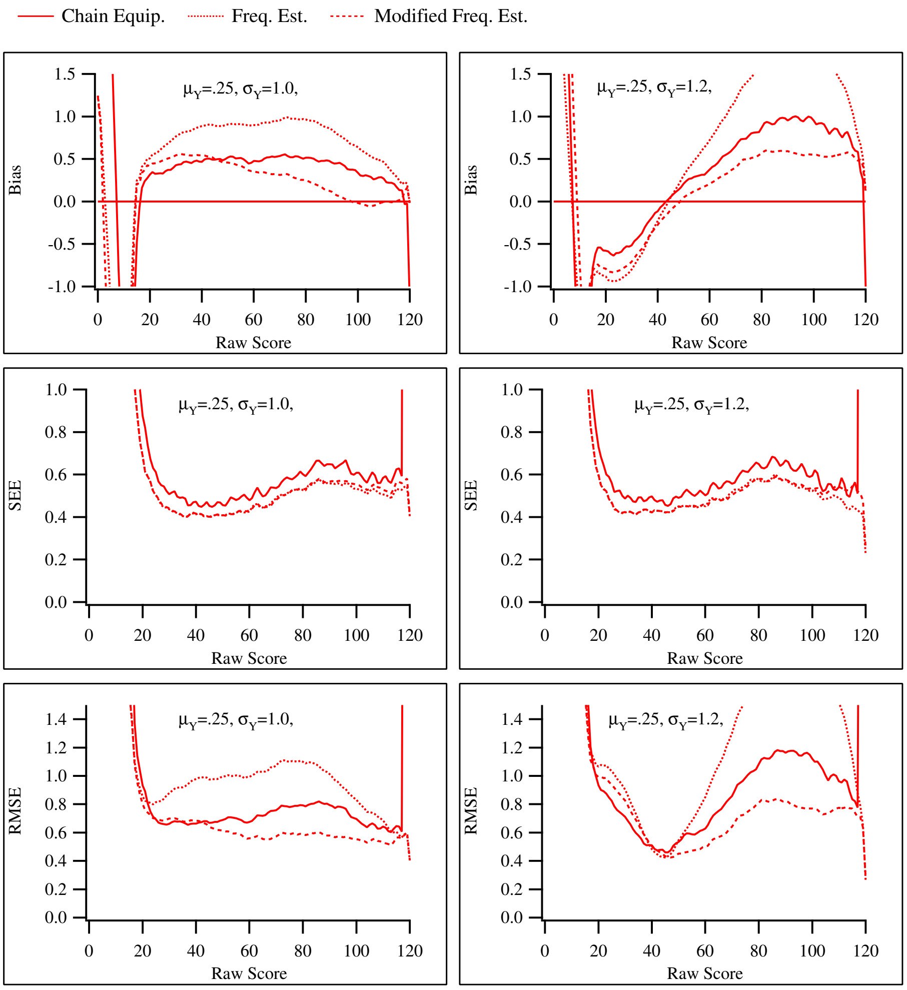

Figure 6: Bias, SEE and RMSE for the 120-Item Test Form Pair with 24-Item Anchor and Unequal Group Variances

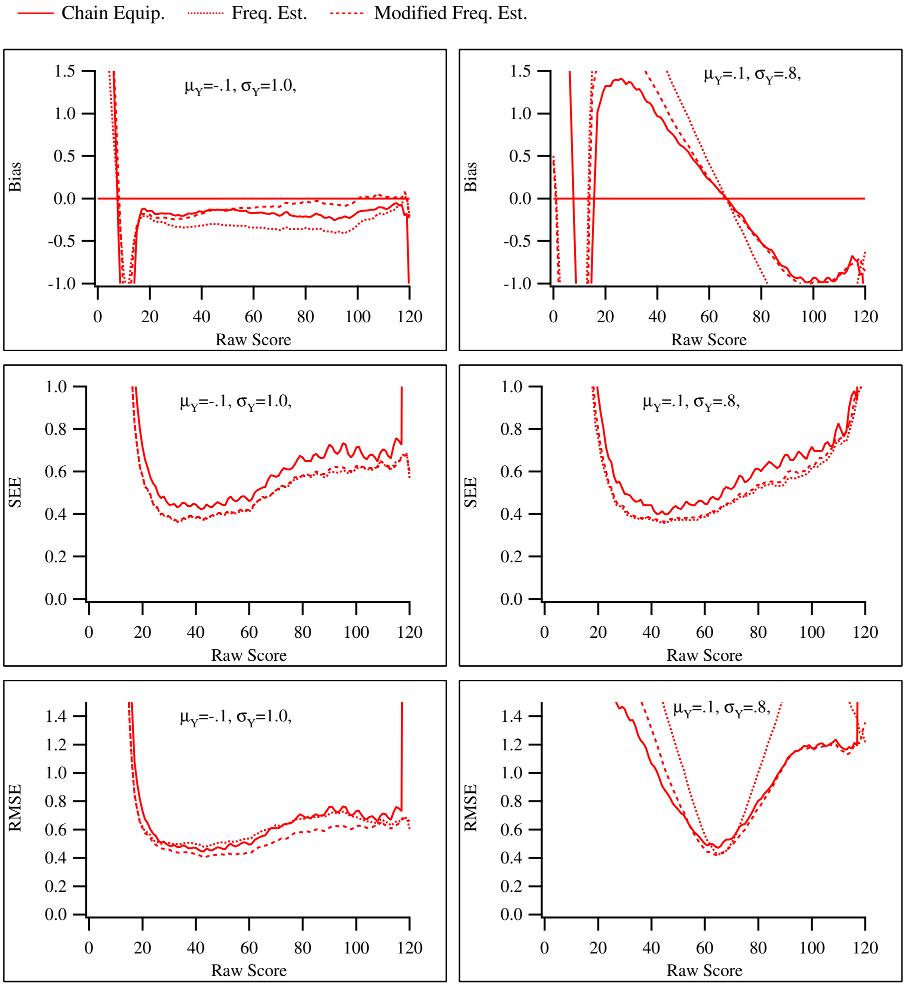

Figure 7: Bias, SEE and RMSE for the 120-Item Test Form Pair with 24-Item Anchor and Group Differences in Opposite Direction