Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 17

A Comparison of the Frequency Estimation and Chained Equipercentile Methods Under the Common-Item Non-Equivalent Groups Design[*]

Tianyou Wang
Won-Chan Lee
Robert L. Brennan
Michael J. Kolen [†]

April 2006

[*] A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April, 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). Won-Chan Lee is Research Scientist, CASMA, University of Iowa. Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, CASMA, University of Iowa. Michael J. Kolen is Professor, College of Education, 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

This paper used simulation to compare two test equating methods under the common-item nonequivalent groups design: the frequency estimation method and the chained equipercentile method. An IRT model was used to define the "true" equating criterion, simulate group differences, and generate response data. Three linear equating methods were also included for reference. The results show that when there is substantial group difference, the frequency estimation method has larger bias than the chained equipercentile method. The frequency estimation method, however, almost always has smaller standard error of equating than the chained equipercentile method.

1 Introduction

In testing equating, there are three commonly used data collection designs: the single-group design with/without counter-balancing, the random groups design, and the common-item nonequivalent groups design (Kolen & Brennan, 2004). Under each design, different equating methods can be considered. When a common-item nonequivalent groups design is used to collect equating data, Form X and Form Y are administered to samples from two different populations. A common set of test items is administered to both samples. The common-item nonequivalent groups design has been used in many important testing programs. Under the common-item nonequivalent groups design, equipercentile equating can be applied using two different methods. The first method is called frequency estimation. For this method, the frequency distributions of Form X and Form Y for a common synthetic population are estimated. Equipercentile equating is then applied to the estimated frequency distributions. The second method is called chained equipercentile equating. For this method, Form X scores are first equated to the common-item V scores in population 1 using the equipercentile equating method. Then, the common-item set V scores are equated to Form Y scores in population 2. Finally, the Form X scores are equated to Form Y scores through a chain consisting of the two equipercentile equating functions. Both these methods have been used in actual testing programs, although they are seldom used in the same testing program.

A few studies have compared the merits of these two methods (Braun & Holland, 1982; Harris & Kolen, 1990; Livingston, Dorans, & Wright, 1990; Marco, Petersen, & Stewart, 1983). These studies found that these two equating methods generally produced quite different results. The studies suggest that the chained equipercentile method may be a better choice when the two groups differ substantially, although the frequency estimation method appears to have better theoretical standing. However, the way these studies were designed limited their ability to accurately evaluate the systematic and random equating errors associated with these methods, because all the studies used real test data as opposed to simulations to compare the two methods. When real test data are used, there is generally a lack of a clear criterion for evaluating systematic equating error. Also, without repeated sampling, it is not possible to evaluate random equating error. Instead, these studies usually aggregate deviations across score points to get so-called root-mean-squared deviations (errors) or overall bias based on only one sample. These inadequacies of the previous studies prevent them from providing a clear and conclusive comparison of the two methods.

von Davier, Holland, and Thayer (2004) did some theoretical analyses of these two methods and showed that they are both examples of what they termed observed score equating, they entail assumptions that are generally not testable in practice, and the two methods produce essentially identical results under two extreme conditions: (1) the two populations are very similar; or (2) the anchor test is perfectly correlated with both tests. Their theoretical work, however, does not illuminate the comparative nature of the two methods under the realistic condition that there is a group difference.

The relatively large differences between the equating results from these two methods and the fact that both methods are regularly used in major testing programs make it imperative to conduct a thorough and systematic study to compare their merits or demerits. In this paper, a simulation technique is used to evaluate and compare the systematic and random equating errors associated with the two methods.

2 Method

The major challenge of conducting such a study is to find a sound way of defining the "true" equating function that can be used as a criterion to evaluate the equating methods. Traditionally, the "true" equating function is defined as the equipercentile equating for the population. Score distributions for the populations are often obtained by smoothing the sample distributions, and the "true" equating functions are computed based on the smoothed population distributions. Then samples are drawn from the population distributions and the equating methods under study are applied to compute the sample-based equating functions, which are compared with the "true" equating function.

One problem with this approach is that if the equating method used to compute the "true" equating function is one of the equating methods under consideration, presumably this will give an unfair advantage to this particular equating method. In order to avoid this problem, the current study will use an IRT approach to define "true" equating and to generate sample data.

2.1 Data Source

This study uses data from four forms of a 60-item Mathematics test. Randomly equivalent groups of about 3000 examinees per form took the test. The item parameters were estimated using BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996) assuming a three-parameter logistic (3PL) IRT model. Following a procedure used by Hanson and Béguin (2002), the item parameters for the four test forms were put on a common scale. The estimated item parameters are treated as "true" parameters that are used to simulate item responses. Parallel test forms with different test lengths are created from these four test forms. A common-item set is created by replacing some of the items from one test with items in the parallel form.

2.2 Simulation Procedure

After parallel test forms with a common-item set are created and "true" item parameters on the common scale are determined, IRT observed-score equating for a population with a standard normal θ distribution is used to compute the "true" equating function. Because IRT observed-score equating is conceptually the equipercentile equating for the population (see Kolen & Brennan, 2004), the equating function computed using this method provides the reasonably defensible definition of "true" equating as long as the IRT assumptions hold.

To simulate a common-item nonequivalent groups design, different population θ distributions are used for the two test forms to be equated. For simplicity, the population θ distribution for the new form (Form X) is always fixed as a standard normal distribution N(0, 1). The population θ distribution for the old form (Form Y) is normal distribution with varying means and standard deviations (SD). The simulation and analyses involve the following steps:

> *Step 1:* Given a pair of population θ distributions, random samples of 2000 θ values are drawn from each of these two θ distributions.
>
> *Step 2:* The response vector of a simulee with a θ value on the corresponding test form is generated based on the item parameters and θ using the IRT model.
>
> *Step 3:* Given simulees' test data for forms X and Y, the frequency estimation and chained equipercentile method are used to compute the sample equating functions.
>
> *Step 4:* Steps 1 to 3 are repeated 500 times. Based on the repeated samples, bias, standard errors of equating (SEE), and root mean squared error (RMSE) are computed and compared.
>
> *Step 5:* Steps 1 to 4 are repeated for different pairs of population θ distributions.
>
> *Step 6:* Steps 1 to 5 are repeated for other conditions included in the study, such as total test length and the ratio of number of common items to total test length.

The advantage of this IRT-based simulation approach is that we can define a reasonable "true" equating function that does not obviously favor either of the equating methods under consideration, and the degree of group differences in the common-item nonequivalent groups design can be easily manipulated. The shortcoming is that we make the assumption that the IRT model is true and thus limit the conclusions of the study to situations in which IRT model fit can be reasonably assumed.

2.3 Factors Studied and Evaluating Indices

This simulation study includes the factors described below:

> *Factor 1:* Test equating methods (frequency estimation and chained equipercentile equating). For reference purposes, we will also include three linear methods: the Tucker method, the Braun-Holland method (Braun & Holland, 1982), and the chained linear equating method. The Braun-Holland method is similar to the frequency estimation method except at the final step. The chained linear method is similar to the chained equipercentile method except at the final

step. The Tucker method is included because it is the most commonly used linear method for this equating design.

Factor 2: The degree of group differences which is measured by the magnitude of differences in the mean and standard deviation (SD) of the two θ distributions from which random samples of θ are drawn. As mentioned before, the population θ distribution for the new form (Form X) is fixed as a standard normal distribution N(0, 1). The population θ distribution for the old form (Form Y) has three variations in mean (.05, .1, .25) and two variations in SD (1.0, 1.2). In test equating, mean differences betweem .05 and .1 are generally considered relatively large, while a mean difference of .25 is usually considered a very large difference. So Form Y has a total of six normal distributions: (1) N(0.05, 1.0), (2) N(0.1, 1.0), (3) N(.25, 1.0), (4) N(.05, 1.2), (5) N(.1, 1.2), and (6) N(.25, 1.2).

Factor 3: This study includes two test lengths: 60 items and 120 items. With the 60-item test length, the first two of the four test forms are used as the X and Y forms. With the 120-item test length, the first two of the four test forms are combined to create form X and and second two forms are combined to create the Y form. The formation of the common item sets is described below.

Factor 4: The ratio of the number of common items to total test length. Two different ratios are included: (1/3 and 1/5) for test length of 60 items. For test length of 120 items, only one ratio (1/5) is included. With the ratio (1/3), the common item set is created by replacing every third item in Form Y with every third item in Form X. With the ratio (1/5), the common item set is created by replacing every fifth item in Form Y with every fifth item in Form X. The starting position of the common-item set is adjusted so that the mean difficulty of the common set is somewhere between the mean difficulties of Form X and Form Y. In this study, we only use an internal anchor, which means the common items are included in the total test.

Given the above factors, we have three pairs of test forms and a total of 18 (3 pairs of test forms by 6 degrees of group differences) simulation conditions.

We also include three additional distributions for Form Y: N(0.0, 1.0), N(-0.1, 1.0), and N(0.1, 0.8). These three distributions are not crossed with all three pairs of parallel forms, but only for the pair with test length of 60 and a ratio of (1/3). The purposes for including N(0.0, 1.0) is to provide a baseline comparison. The purposes for including N(-0.1, 1.0) is to compare it with N(0.1, 1.0) and examine the effects of having mean differences in different directions. The purposes for including N(0.1, 0.8) is to compare it with N(0.1, 1.2) and examine the effects of having SD differences in different directions. Altogether, we have 21 simulation conditions.

The descriptive statistics for the common sets and total tests are in Tables 1, 2, and 3. The parameter distributions for the three pairs of test forms are quite typical of those of real world test forms. In particular, the mean difficulty levels of the common-item sets are always between those of Form X and Form Y. This helps ensure that the common-item sets are representative of the test forms, at least in terms of item difficulty. Note that in Tables 1 and 2, the descriptive statistics for Form X are the same while those for the common-item sets and Form Y are different. That is because the common-item sets are always taken from Form X and an internal anchor is used, so Form X remains unchanged. Form Y is changed because part of the items are replaced by the common-item set.

The evaluation indices are: (1) Conditional bias, standard error of equating (SEE), root mean squared error (RMSE); and (2) aggregate bias, SEE, and RMSE. Conditional bias is defined as the difference between mean equated scores across 500 replications and the true equating function conditioned on a particular test score. The conditional SEE is the standard deviation of the equated scores across 500 replications. RMSE is computed from bias and SEE by the following equation:

$$RMSE^2 = Bias^2 + SEE^2. \tag{1}$$

Aggregate error indices are the weighted average of conditional error indices across score points, weighted by the expected probability distribution of Form X, which is used in computing the IRT observed equating function in order to determine the "true" equating function. Note the aggregate bias is the weighted average of the absolute value of conditional bias because otherwise if there are both positive and negative bias along the score scale, they cancel each other out in the summation and give a false impression that there is not much bias.

3 Results

The equating bias and SEE of the four equating methods conditional on each score point are plotted in Figures 1 through 8. Figure 1 is the baseline condition with Form Y having a distribution of N(0.0, 1.0). Figures 2 and 3 are for the pair of 60-item test forms with 20 common items. Figures 4 and 5 are for the pair of 60-item test forms with 12 common items. Figures 6 and 7 are for the pair of 120-item test forms with 24 common items. Figures 8 is for the conditions where Form Y has a distribution of N(-0.1, 1.0) and N(0.1, 0.8), respectively. We omitted some of the plots in this paper. A more complete set of plots are in Wang, Lee, Brennan and Kolen (2006).

3.1 General Comparisons Among the Equating Methods

Figure 1 shows that when there is no group difference, both the frequency estimation method and the chained equipercentile method are essentially unbiased except at the very lower end of the scale. Figures 2 through 7 show that if there is a group mean difference, the magnitude of the bias for the frequency estimation method is always larger than for the chained equipercentile method. By contrast, the SEE for the frequency estimation method is always lower than for the chained equipercentile method. The SEE's for both methods seem little affected by group differences. The comparisons of the overall equating error for the two methods as reflected in the RMSE plots, depend largely on group difference. When group differences are small, frequency estimation has less RMSE; when group differences increase, chained equipercentile method tends to show less RMSE.

The linear methods display almost exactly the same bias when there is no group difference as shown in Figure 1. When group mean differences increase, the linear methods tend to have differential bias but the shape of the bias curves remain the same. The shape of the bias curve for the linear methods basically reflect the discrepancy between the "true" equating relationship and a linear equating relationship. There is always a bias for the linear methods as long as the true equating function is not linear. The Tucker method has slightly larger bias at the upper end of the score scale than the other two linear methods, and the Braun-Holland method has larger bias at the lower end of the score scale than the other two linear methods. The chained linear method most often gives the least bias of the three linear methods.

The linear methods always have smaller SEE's than the two equipercentile methods, just as expected. Among the linear methods, the Braun-Holland method tends to have slightly smaller SEE than the other two methods, especially towards both ends of the score scale. The differential magnitudes of RMSE between the linear methods and the equipercentile methods depend primarily on the bias comparison (which in turn depends on the the nonlinearity of the "true" equating relationship, as mentioned above). The lower bias for the linear methods translates into relatively smaller RMSE for the linear methods than for the equipercentile methods.

3.2 The Effect of Group Difference on Equating Error

Comparisons among Figures 2 through 7 clearly show that larger group mean difference tend to produce more bias and tend to enlarge the bias difference between frequency estimation and chained equipercentile methods. Figures 2 and 8 show that if the group means differ in a different direction, then the bias shift direction, but the shape of the bias curves remain unchanged.

Comparisons between plots in Figures 3, 5 and 7 ($\sigma_Y = 1.0$ versus $\sigma_Y = 1.2$) show that differences in group variability affect the slope of the bias curve. When the new group variability is smaller, it produces a negative bias on the lower side of the score scale and positive bias on the upper side of the score scale. This relationship is reversed in Figure 8 where the new group has larger variability.

The bias pattern in Figures 1 through 8 suggests that group difference is the main source of bias for the two equipercentile methods, rather than some inherent bias embedded in the methods as in the linear case. However, the chained equipercentile method seems less susceptible to group mean difference in terms of bias than the frequency estimation method.

3.3 The Effect of Ratio of Number of Common Items to Total Test Length on Equating Error

Comparisons between Figures 2, 3 and Figures 4, 5 show that a smaller ratio of common items tends to increase both the bias and SEE for all the methods. Note that the differences in the bias pattern of the linear methods between these two sets of plots is mainly due to the different non-linearity of the true equating functions of these two pairs of test forms, and probably less due to the change in the ratio.

3.4 The Effect of Total Test Length on Equating Error

Comparisons between Figures 4, 5 and Figures 6, 7 show that equating errors tend to increase as total test length increases. This may be due to the fact that we used the same sample size (2000) for all the simulations. Having more items means more score categories, and the frequency of each category decreases. This may contribute to larger bias and SEE. This is consistent with the common knowledge that longer test length requires a larger sample size to maintain the same equating precision.

3.5 Overall Equating Errors

Results for the aggregate equating errors are in Tables 4, 5, and 6. The smallest errors among the linear methods or equipercentile methods are highlighted. These results tend to confirm the general patterns found in the plots for conditional equating errors. The frequency estimation method tends to have larger bias than the chained equipercentile method when there are group differences. The frequency estimation method has smaller SEE's than the chained equipercentile method. The SEE's for all the methods are essentially unaffected by group differences. The RMSE comparison depends on how large the group differences are. The larger the group differences, the more likely the chained equipercentile methods performs better in terms of lower RMSE. Specifically, for this simulation RMSE for chained equipercentile was always less than that for frequency estimation when SD differed and/or mean difference were greater than 0.1. The linear methods always have significantly less SEE than the equipercentile methods. The Braun-Holland method performs the best when there is small group differences. When group differences increase, the chain linear method tends to perform better than the Braun-Holland method. A strikingly consistent result is that the Tucker method never performs the best among the linear methods. The bias and RMSE comparisons between the linear and equipercentile methods are mixed. Among the linear methods, the chained linear methods seems to produce better results than the other linear mehtods, especially when there are large group differences.

4 Conclusions and Discussion

Previous literature based on real test data have found that the frequency estimation method and chained equipercentile method produced quite different results. This study used IRT as the psychometric model to establish a "true" equating function and to simulate test data to evaluate these two methods, along with three linear methods. The results show that generally speaking, the frequency estimation method does produce more bias than chained equipercentile method and the difference in bias increases as group differences increase. Based on these results we recommend the frequency estimation method when group differences are small, and the chained equipercentile method when group differences are large. To assess the magnitude of group differences, the mean and SD of the common item set from both populations can be compared.

To explain why the frequency estimation methods consistently have smaller SEE than the chained methods (which is true for both the equipercentile and linear cases), we conjecture that this is probably due to the fact that the frequency estimation methods utilize two bivariate distributions while the chained methods only uses two pairs of marginal distribution. It is expected that estimation methods utilizing more information in the data may result in more stable estimates. This result also suggests that in order to achieve similar SEE's, the frequency estimation methods requires smaller sample size than the chained methods.

As to the reason why frequency estimation method produces more bias than the chained equipercentile method, we conjecture that because the frequency estimation method makes a strong assumption about the equality of conditional distributions in the two populations, when there are substantial group differences, this assumption may be violated. The chained equipercentile method, however, does not make such an overt assumption.

Earlier it was mentioned that the IRT procedure used here for establishing the "true" equating should provide a "reasonably defensible" criterion that does not seriously advantage/disadvantage any of the methods under study. On the one hand, it is virtually impossible in a simulation to have a criterion that does not advantage/disadvantage one or more methods to some degree. In this particular simulation, it is possible that the criterion disadvantages the frequency estimation method to some unknown extent. Specifically, there is nothing in the procedure used to establish the criterion that accommodates or reflects the assumption that, conditional on anchor test scores, Form X scores in the two populations are the same; similarly for Form Y scores. One could conceive of constructing the criterion in such a manner that these conditional distributions are indeed equal. If that were done, it seems reasonable to speculate that frequency estimation might have considerably less bias than exhibited in this study.

Also, in constructing the criterion, a set of items, say V, from Form X were declared to be the common items. Then these items replaced items in Form Y. This means that the item parameters for V in X are identically the same item parameters as for V in Form Y. This makes sense for establishing "true" equating for simulation, but it may not reflect the reality of what happens in operational equating–namely, context effects may influence scores on V differentially in X (for one population) and Y (for the other population). Whether chained equipercentile equating or frequency estimation is more susceptible to such context effect is beyond the scope of this paper, but needs to be considered in practice.

The limited scope of the paper prevents it from including all possible factors. We included only one sample size (2000) and one set of weights (.5 and .5) for forming the synthetic population. We made these choices because they are commonly used in practice. Also, The simulation study in this paper did not employ any smoothing procedure on score distributions. How variations on these factors affect the results requires further empirical study.

Finally, since the simulations in this study were carried out under an IRT framework, the results of this paper apply to situations where an IRT model can be reasonably assumed. How violations of IRT assumptions affect the validity of the results of this paper should be further investigated.

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.

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

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.

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.

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.* Paper presented at the annual meeting of National Council of Measurement in Education, April, San Francisco.

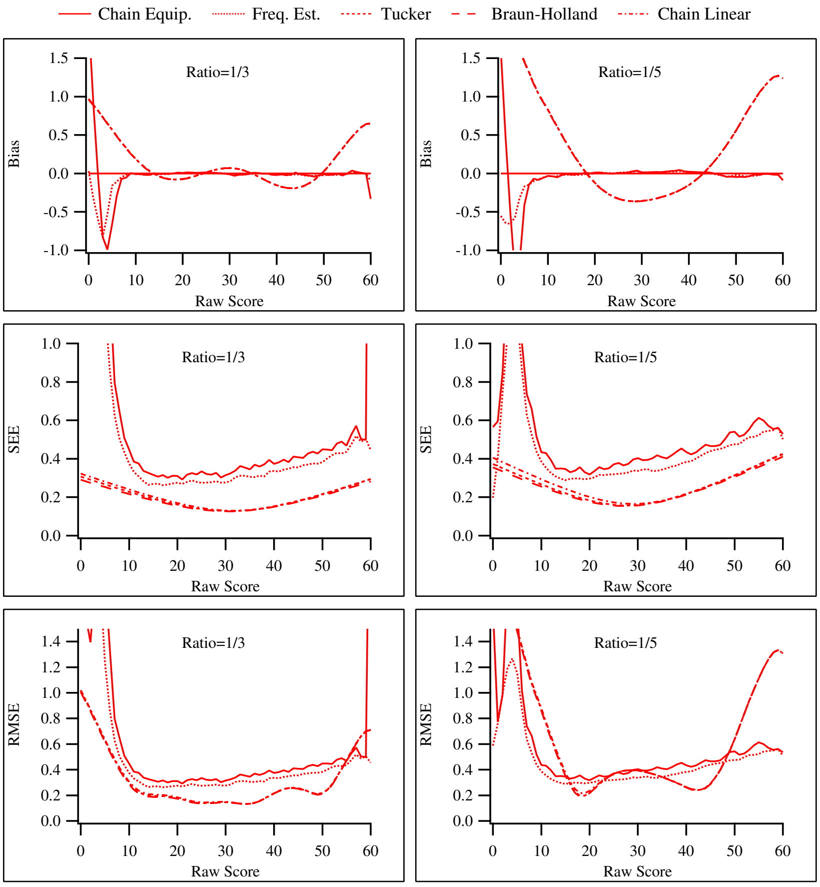

Figure 1: Bias, SEE and RMSE for the Two Pairs of 60-Item Forms When There is No Group Difference(μ_Y=.0, σ_Y=1.0, Length=60)

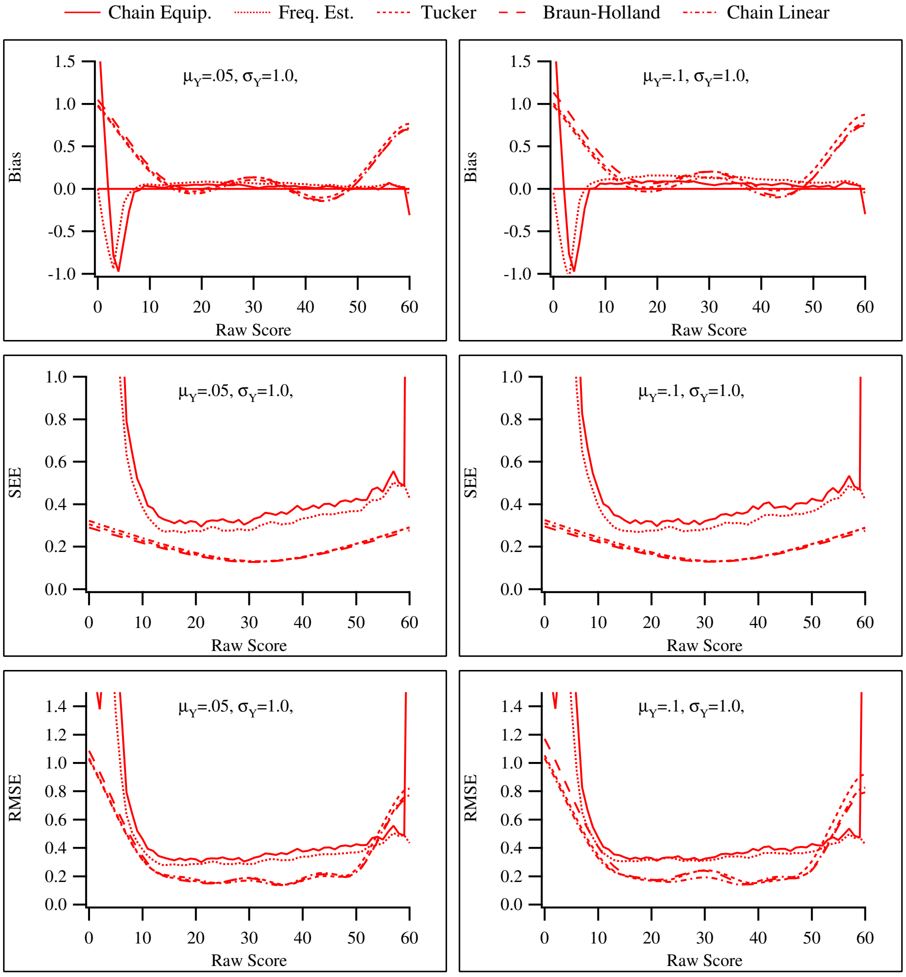

Figure 2: Bias, SEE and RMSE for the First Pair of Test Forms (Length=60, Ratio=1/3) and Equal Group Variances

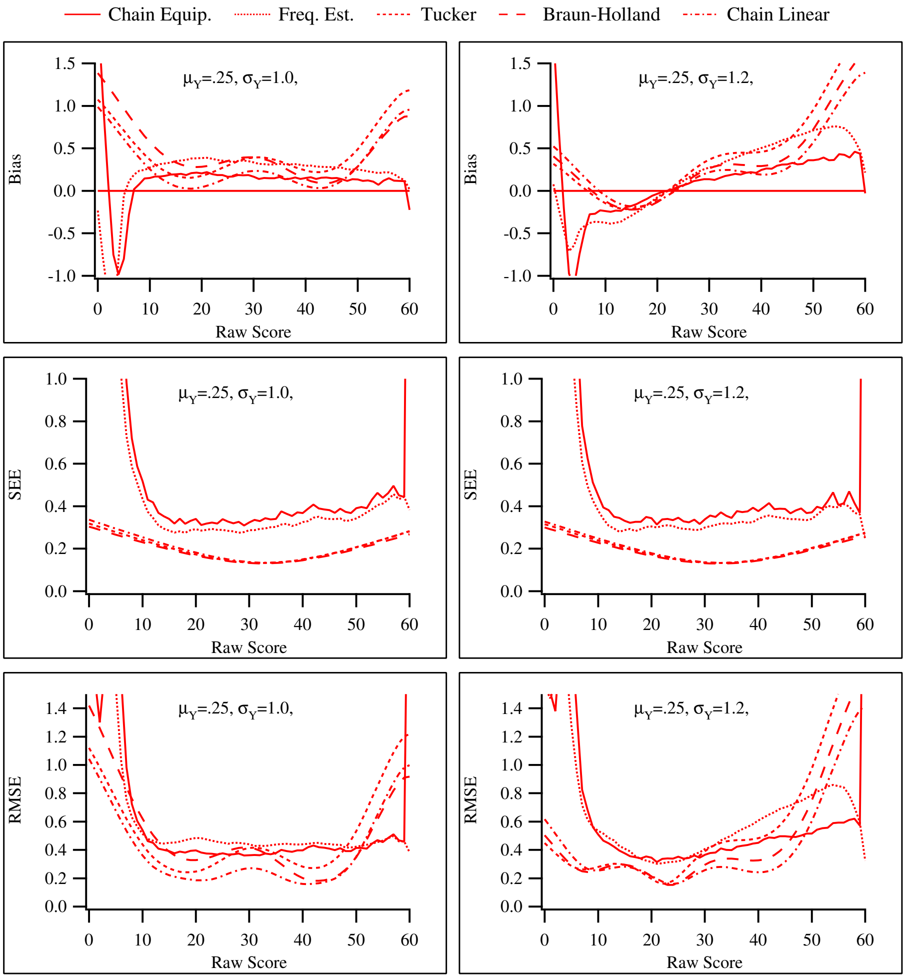

Figure 3: Bias, SEE and RMSE for the First Pair of Test Forms (Length=60, Ratio=1/3) and Different Group Variances

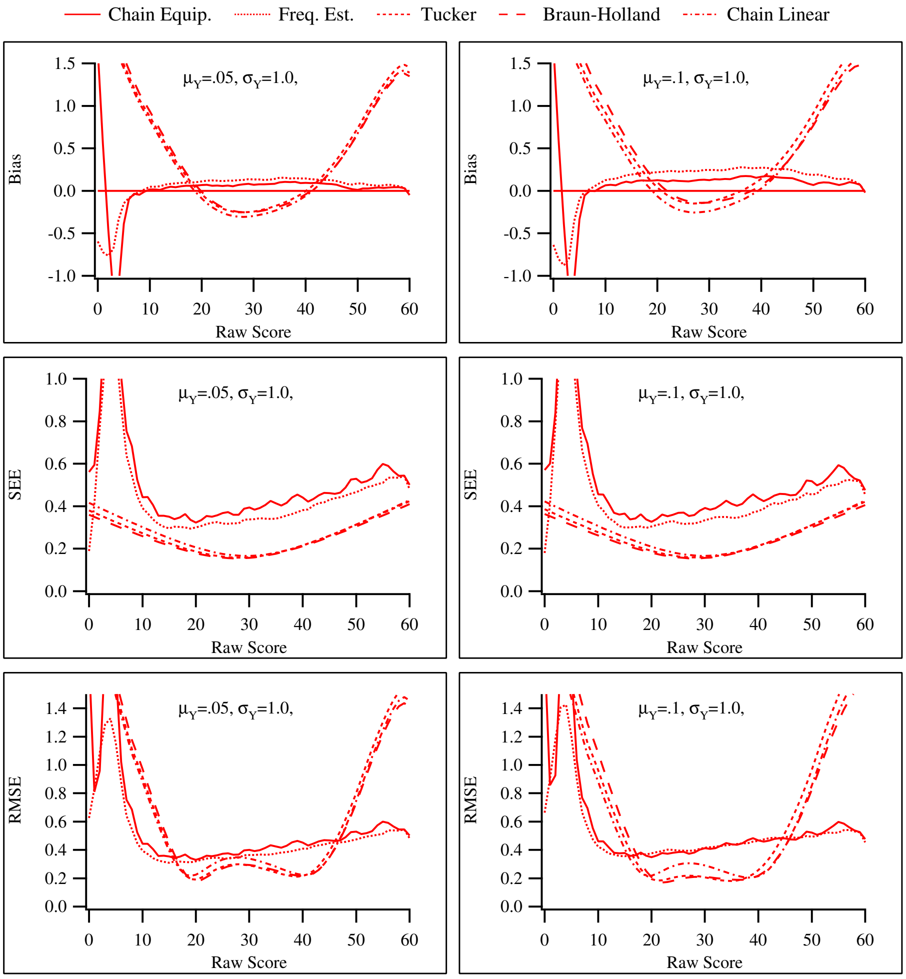

Figure 4: Bias, SEE and RMSE for the Second Pair of Test Forms (Length=60, Ratio=1/5) and Equal Group Variances

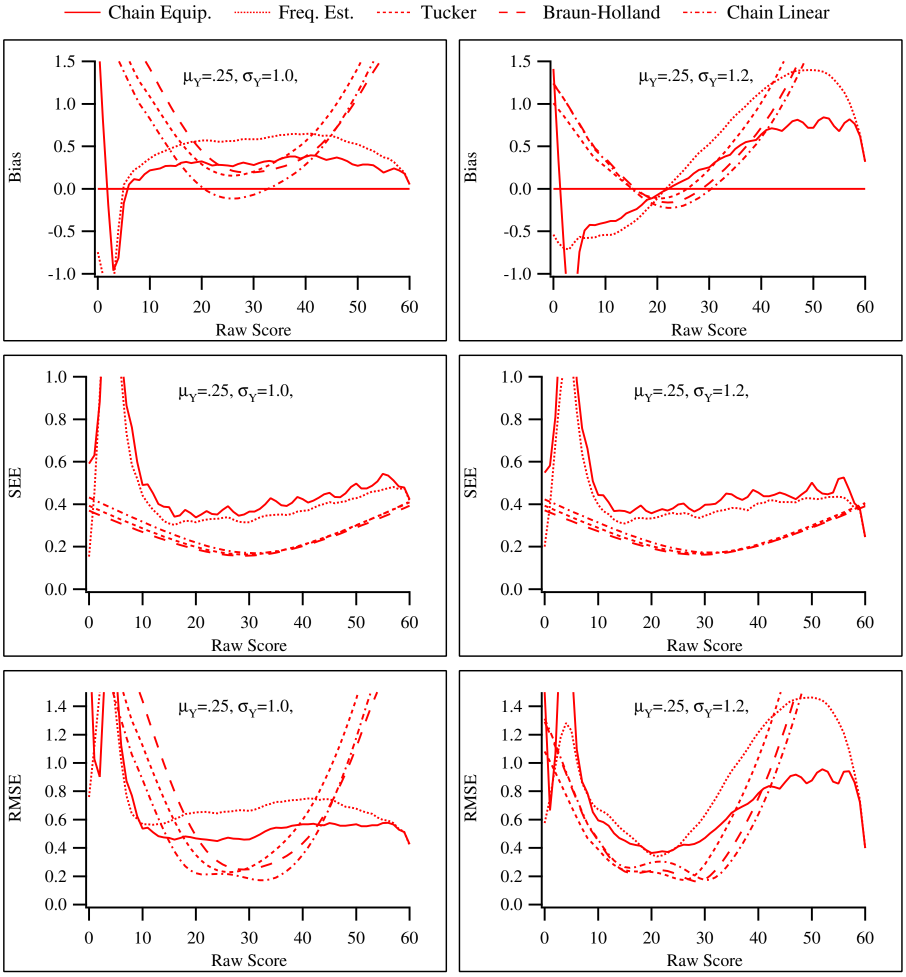

Figure 5: Bias, SEE and RMSE for the Second Pair of Test Forms (Length=60, Ratio=1/5) and Different Group Variances

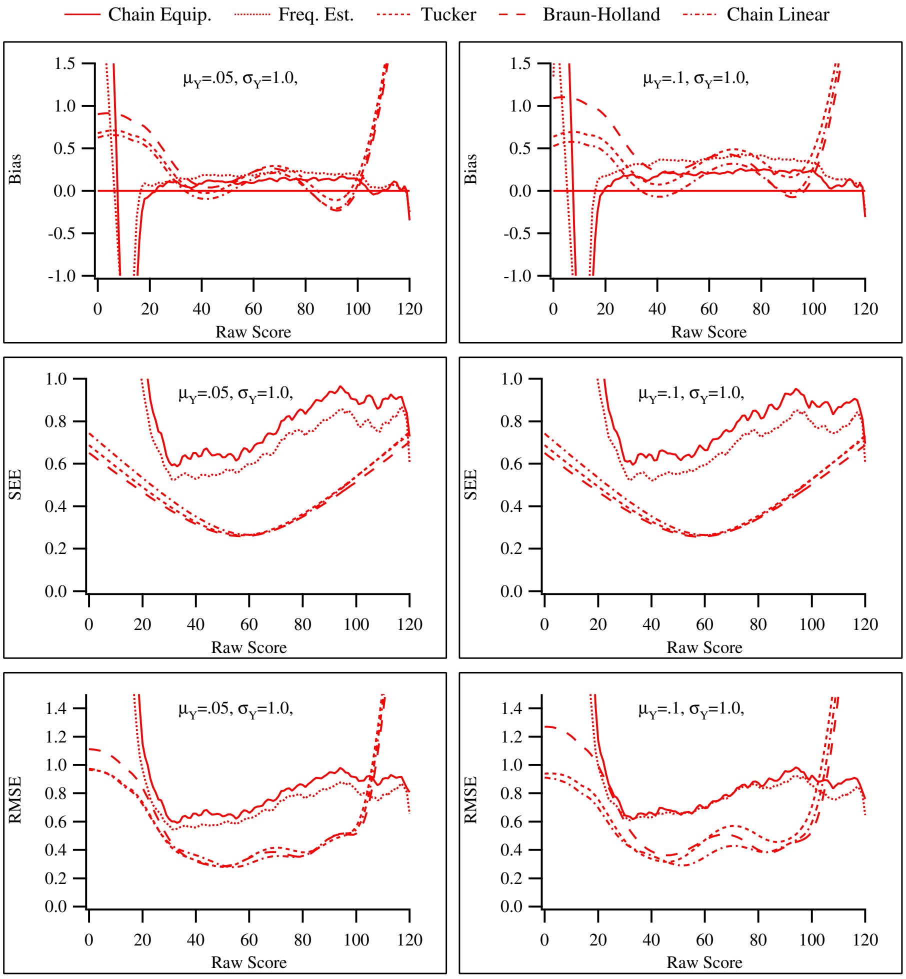

Figure 6: Bias, SEE and RMSE for the Third Pair of Test Forms (Length=120, Ratio=1/5) and Equal Group Variances

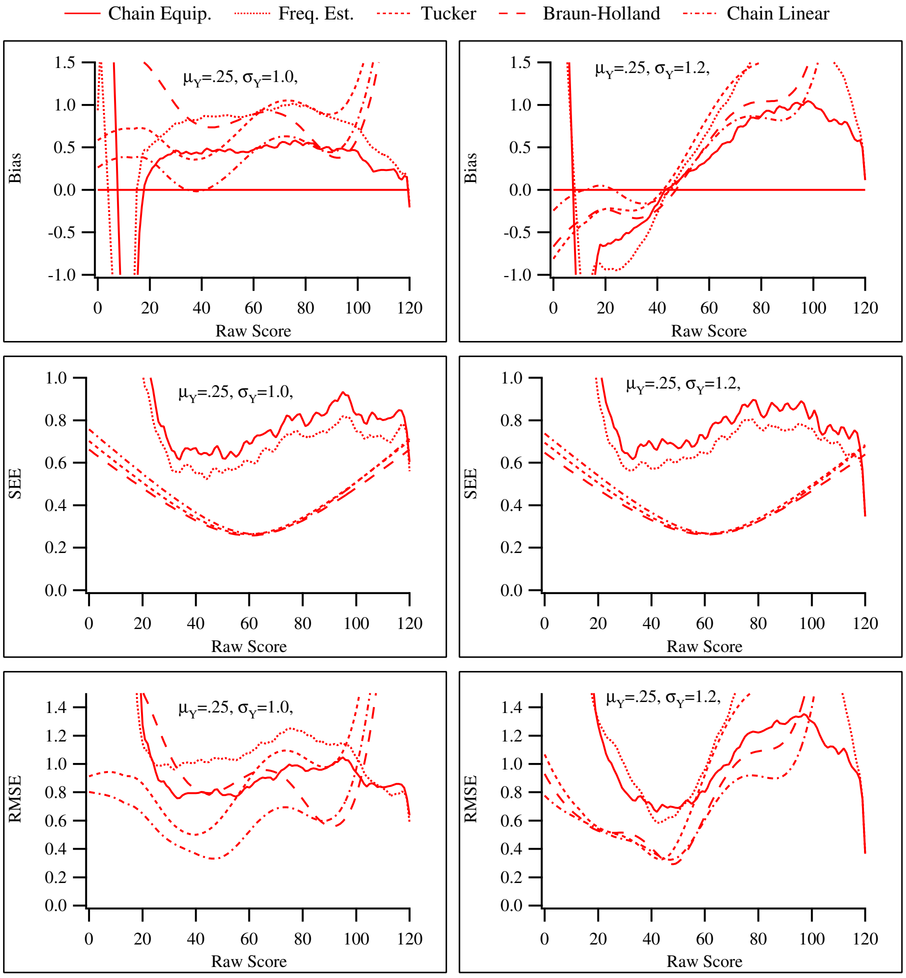

Figure 7: Bias, SEE and RMSE for the Third Pair of Test Forms (Length=120, Ratio=1/5) and Different Group Variances

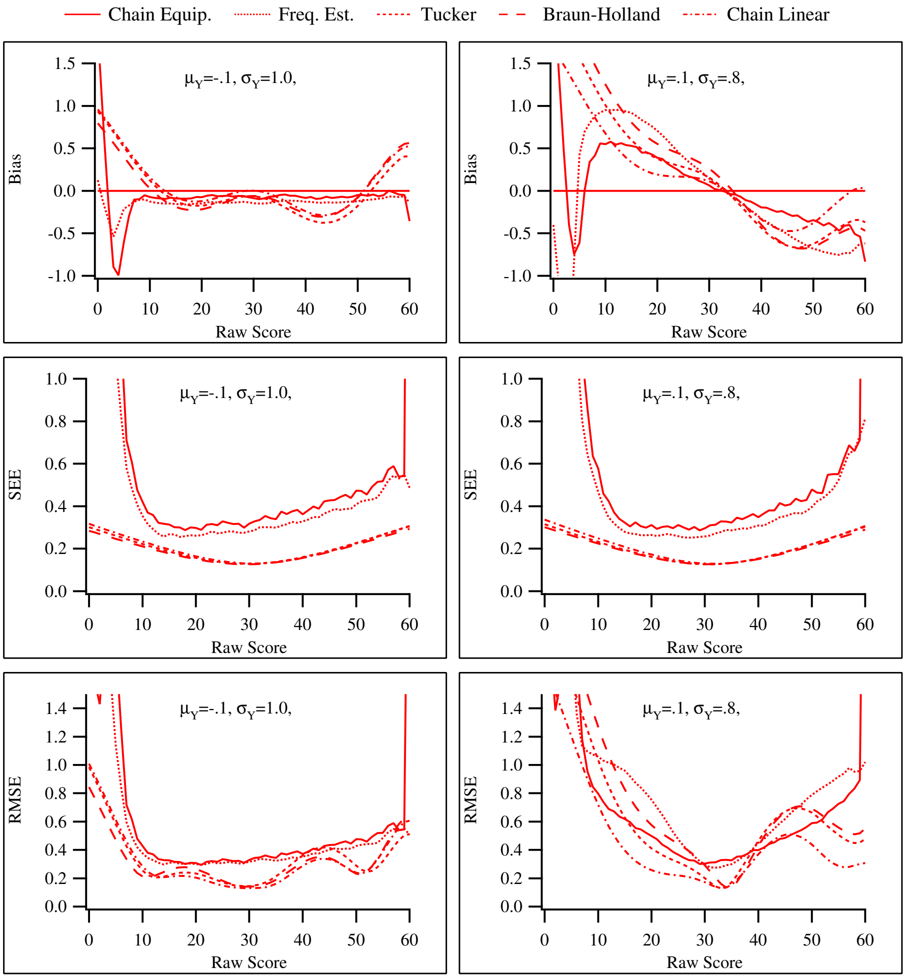

Figure 8: Bias, SEE and RMSE for the First Pair of Test Forms (Length=60, Ratio=1/3) with Mean and Variance Difference in Opposite Direction

Table 1: Descriptive Statistics of Item Parameters For Test Length of 60 and Ratio of 1/3

Parameter	n	Mean	SD	Skewness	Kurtosis
Common-Item Set					
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

Table 2: Descriptive Statistics of Item Parameters For Test Length of 60 and Ratio of 1/5

Parameter	n	Mean	SD	Skewness	Kurtosis
Common-Item Set					
a	12	1.040138	0.331521	0.763516	2.037446
b	12	0.192165	0.961684	-0.016920	2.248588
c	12	0.132799	0.040575	-0.348299	1.742651
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	0.976284	0.318770	0.387069	2.353556
b	60	0.154999	1.015040	-0.327913	2.375470
c	60	0.150104	0.040367	0.094340	2.742320

Table 3: Descriptive Statistics of Item Parameters For Test Length of 120 and Ratio of 1/5

Parameter	n	Mean	SD	Skewness	Kurtosis
Common-Item Set					
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

Table 4: Aggregate Equating Errors For Test Length of 60 and Ratio of 1/3

	Equipercentile		Linear		
Index	Chain Equip.	Freq. Est.	Tucker	Chain Lin.	B-H
$(\mu_Y = 0.0, \sigma_Y = 1.0)$					
Abs. Bias	0.013280	0.010292	0.093254	0.093155	0.093450
SEE	0.365246	0.313869	0.165486	0.169016	0.160514
RMSE	0.365707	0.314144	0.198779	0.202089	0.194547
$(\mu_Y = 0.05, \sigma_Y = 1.0)$					
Abs. Bias	0.032253	0.060970	0.096728	0.091666	0.099301
SEE	0.364353	0.313860	0.165943	0.169754	0.161057
RMSE	0.366256	0.320611	0.204120	0.202808	0.200706
$(\mu_Y = 0.1, \sigma_Y = 1.0)$					
Abs. Bias	0.065229	0.125692	0.130567	0.096956	0.141365
SEE	0.363385	0.314139	0.167187	0.170988	0.162338
RMSE	0.370152	0.340583	0.228510	0.208987	0.225687
$(\mu_Y = 0.25, \sigma_Y = 1.0)$					
Abs. Bias	0.166583	0.320030	0.317809	0.159647	0.318291
SEE	0.366330	0.315625	0.169758	0.173857	0.164562
RMSE	0.405036	0.454302	0.366578	0.251842	0.363811
$(\mu_Y = 0.05, \sigma_Y = 1.2)$					
Abs. Bias	0.165279	0.312154	0.294746	0.182655	0.283824
SEE	0.370198	0.320490	0.166413	0.168771	0.160876
RMSE	0.416688	0.466631	0.348911	0.260835	0.338328
$(\mu_Y = 0.1, \sigma_Y = 1.2)$					
Abs. Bias	0.166124	0.312605	0.316710	0.195082	0.282430
SEE	0.367741	0.320317	0.166348	0.168698	0.160675
RMSE	0.413788	0.465644	0.369692	0.270193	0.335306
$(\mu_Y = 0.25, \sigma_Y = 1.2)$					
Abs. Bias	0.178853	0.338050	0.386507	0.231685	0.294121
SEE	0.368165	0.320865	0.168834	0.171106	0.163175
RMSE	0.418784	0.485298	0.441786	0.303808	0.351922
$(\mu_Y = -0.1, \sigma_Y = 1.0)$					
Abs. Bias	0.072296	0.135481	0.170593	0.117147	0.170226
SEE	0.363314	0.313343	0.166779	0.169540	0.161469
RMSE	0.371148	0.343085	0.247172	0.217391	0.241170
$(\mu_Y = 0.1, \sigma_Y = 0.8)$					
Abs. Bias	0.268614	0.486309	0.401818	0.249222	0.495579
SEE	0.365752	0.312809	0.166549	0.171979	0.163219
RMSE	0.473522	0.602552	0.442363	0.312574	0.528768

Table 5: Aggregate Equating Errors For Test Length of 60 and Ratio of 1/5

	Equipercentile		Linear		
Index	Chain Equip.	Freq. Est.	Tucker	Chain Lin.	B-H
($\mu_Y = 0.05, \sigma_Y = 1.0$)					
Abs. Bias	0.065965	0.113202	0.305717	0.312942	0.303402
SEE	0.407970	0.358788	0.212363	0.222366	0.205988
RMSE	0.414121	0.377931	0.391912	0.402109	0.385895
($\mu_Y = 0.1, \sigma_Y = 1.0$)					
Abs. Bias	0.124937	0.223200	0.313216	0.308703	0.303881
SEE	0.407693	0.359142	0.213059	0.223151	0.205636
RMSE	0.427868	0.426306	0.405382	0.401958	0.393294
($\mu_Y = 0.25, \sigma_Y = 1.0$)					
Abs. Bias	0.306017	0.556810	0.566362	0.346188	0.552174
SEE	0.404254	0.358556	0.215059	0.225317	0.205829
RMSE	0.510542	0.667855	0.616074	0.446716	0.596581
($\mu_Y = 0.05, \sigma_Y = 1.2$)					
Abs. Bias	0.325496	0.557084	0.530193	0.417176	0.546326
SEE	0.419825	0.372885	0.216008	0.222463	0.209943
RMSE	0.551621	0.696846	0.599257	0.500314	0.604930
($\mu_Y = 0.1, \sigma_Y = 1.2$)					
Abs. Bias	0.332597	0.571643	0.547742	0.428238	0.521575
SEE	0.416992	0.372169	0.216242	0.222706	0.209756
RMSE	0.553387	0.708375	0.619816	0.512614	0.585541
($\mu_Y = 0.25, \sigma_Y = 1.2$)					
Abs. Bias	0.372829	0.657024	0.646704	0.476618	0.514363
SEE	0.413140	0.369868	0.217953	0.224628	0.209338
RMSE	0.582890	0.786417	0.731651	0.566268	0.594321

Table 6: Aggregate Equating Errors For Test Length of 120 and Ratio of 1/5

	Equipercentile		Linear		
Index	Chain Equip.	Freq. Est.	Tucker	Chain Lin.	B-H
$(\mu_Y = 0.05, \sigma_Y = 1.0)$					
Abs. Bias	0.107881	0.178482	0.208374	0.181217	0.225280
SEE	0.752150	0.659034	0.360358	0.368694	0.347937
RMSE	0.760825	0.684685	0.449243	0.437005	0.438540
$(\mu_Y = 0.1, \sigma_Y = 1.0)$					
Abs. Bias	0.198026	0.346673	0.359892	0.219143	0.354643
SEE	0.751159	0.657653	0.357233	0.365116	0.343938
RMSE	0.778678	0.748141	0.534374	0.462861	0.519162
$(\mu_Y = 0.25, \sigma_Y = 1.0)$					
Abs. Bias	0.461143	0.843457	0.844900	0.440278	0.840018
SEE	0.749902	0.660271	0.356369	0.364667	0.341732
RMSE	0.885709	1.081249	0.930886	0.621965	0.915048
$(\mu_Y = 0.05, \sigma_Y = 1.2)$					
Abs. Bias	0.448330	0.802022	0.689336	0.404594	0.704771
SEE	0.762122	0.675376	0.353846	0.357344	0.342304
RMSE	0.910505	1.089948	0.801417	0.578835	0.801376
$(\mu_Y = 0.1, \sigma_Y = 1.2)$					
Abs. Bias	0.461256	0.820850	0.769320	0.460717	0.703393
SEE	0.761092	0.676258	0.351962	0.355903	0.340581
RMSE	0.916671	1.105021	0.878962	0.626810	0.799567
$(\mu_Y = 0.25, \sigma_Y = 1.2)$					
Abs. Bias	0.530078	0.942869	1.010095	0.609320	0.749125
SEE	0.765799	0.680338	0.354395	0.358932	0.340396
RMSE	0.959826	1.208297	1.123647	0.765421	0.862048