Research Report

Kernel and Traditional Equipercentile Equating With Degrees of Presmoothing

Tim Moses

Paul Holland

Research & Development

April 2007
RR-07-15

Kernel and Traditional Equipercentile Equating With Degrees of Presmoothing

Tim Moses and Paul Holland

ETS, Princeton, NJ

April 2007

Kernel and Traditional Equipercentile Equating With Degrees of Presmoothing

Tim Moses and Paul Holland

ETS, Princeton, NJ

April 2007

Abstract

The purpose of this study was to empirically evaluate the impact of loglinear presmoothing accuracy on equating bias and variability across chained and post-stratification equating methods, kernel and percentile-rank continuization methods, and sample sizes. The results of evaluating presmoothing on equating accuracy generally agreed with those of previous presmoothing studies, suggesting that less parameterized presmoothing models are more biased and less variable than highly parameterized presmoothing models and raw data. Estimates of standard errors of equating were most accurate when based on large sample sizes and score-level data that were not sparse. The accuracy of standard error estimates was not influenced by the correctness of the presmoothing model. The accuracy of estimates of the standard errors of equating differences was also evaluated. The study concludes with some detailed comparisons of how the kernel and traditional equipercentile continuization methods interacted with data that were presmoothed to different degrees.

Key words: Equating, NEAT design, kernel, equipercentile, loglinear presmoothing

Table of Contents

Page

List of Tables

List of Figures

Page

Introduction

The equating literature has considered the implications of loglinear presmoothing (Holland & Thayer, 1987, 2000) on the accuracy of equipercentile equating functions (i.e., the bias and variability of equated scores) and the accuracy of estimated standard errors of equating (SEE). Prior evaluations of presmoothing have been across equating designs, equating methods, and continuization methods for estimating "in-between" scores in discrete distributions; and in reference to particular population distributions. These prior studies, reviewed next, have focused on either equating function accuracy or on SEE accuracy. This study extends previous considerations of presmoothing on equating function and SEE accuracy in the nonequivalent groups with anchor test (NEAT) equating design (i.e., where one of two test forms, X and Y, and an anchor test, A, are given to independent samples of one of two different populations, P and Q, and X_P is equated to test Y_Q using anchor scores A_P and A_Q to account for ability differences in the populations).

Presmoothing and Equating Function Accuracy

Prior studies of loglinear presmoothing on equating function accuracy are distinguished in terms of the equating designs they considered and their defined criterion equating function. Livingston (1993) evaluated presmoothing and NEAT chained equipercentile equating where the criterion equating function was an available single-group equipercentile function based on unsmoothed test data. Hanson, Zeng, and Colton (1994) evaluated presmoothing on equivalent groups equipercentile equating functions where the criterion equivalent-groups functions were based on presmoothed distributions from known loglinear or beta-binomial models, or on unsmoothed data. Hanson (1991) evaluated bivariate presmoothing for frequency estimation equipercentile equating where the criterion equating functions were frequency estimation functions based on presmoothed distributions from loglinear and beta-binomial models. In these studies, the traditional equipercentile method based on the percentile-rank continuization method (Kolen & Brennan, 2004) was exclusively considered. The general result in these studies was that highly parameterized loglinear presmoothing models produced less biased and more variable equating functions than did less parameterized models.

Presmoothing and SEE Accuracy

Prior studies of loglinear presmoothing on SEE accuracy have focused on comparing SEEs based on raw and smoothed data across small and large sample sizes. Liou and Cheng (1995) compared SEEs based on raw and smoothed data for chained, frequency estimation, and single-group equipercentile equating functions across small (100 and 200) and large (1,000 and 2,000) sample sizes, where the smoothed SEEs were estimated using the population loglinear models used to generate the sample data. Liou, Cheng, and Johnson (1997) compared SEEs based on raw and smoothed data for frequency estimation and chained equipercentile functions across percentile-rank and kernel (Holland & Thayer, 1989; von Davier, Holland, & Thayer, 2004) continuization methods and small (100) and large (1,000) sample sizes. Liou et al. (1997) considered population distributions from known loglinear models and also from three-parameter IRT models. The results of these studies were that SEE accuracy was strongly affected by sample size and was always higher for smoothed data than for raw data, regardless of whether the presmoothing model was correct.

This Study

The focus of this study is on the impact of the loglinear presmoothing model on equating function accuracy and SEE accuracy. This evaluation is across sample size, NEAT equating method (chained and post-stratification/frequency estimation equipercentile) and continuization method (kernel and percentile-rank). This study builds on the results of prior studies in several ways:

- The previous studies suggest that the correctness of the presmoothing model affects equating function bias but *not* SEE accuracy. This suggestion is a direct focus of the current study.
- This study evaluates a possible interaction between the kernel continuization method and the degree of presmoothing. Liou et al. (1997) considered the kernel continuization method but did so using fixed kernel bandwidths (i.e., degrees of continuization). The more current proposal for implementing kernel equating (von Davier et al., 2004) is to use kernel bandwidths that vary according to the sample data, meaning that the kernel continuization functions like a post-smoothing method when presmoothed data are rough. It is likely that sample-dependent kernel bandwidths have implications for the accuracy of equating functions and SEEs that depend on whether the data are strongly

or weakly presmoothed. In addition, the implications of kernel bandwidths across degrees of presmoothing should differ from the implications of using the sample-independent percentile-rank continuization method.

- The two major NEAT equating methods, chained and post-stratification, may respond differently to varied degrees of presmoothing. Each method utilizes presmoothed data in different ways. Where chained equating focuses solely on the univariate smoothing results (i.e., the marginal X_P, A_P, A_Q, and Y_Q distributions), post-stratification equating focuses on the bivariate smoothing results (i.e., the *XA* distribution in *P* and the *YA* distribution in *Q*). It is possible that the equating methods' different uses of presmoothing may result in unique implications for equating function accuracy and SEE accuracy.

  This study's consideration of chained and post-stratification equating methods is very different from the consideration of which equating method is more biased and variable with respect to a true equating function (e.g., Sinharay & Holland, 2006; Wang, Lee, Brennan, & Kolen, 2006). The concern of this study is the statistical bias from incorrectly estimating the population loglinear presmoothing model, rather than the equating method bias from dealing incorrectly with the missing data in the NEAT design. To directly evaluate the impact of the correctness of the loglinear presmoothing model across equating methods, the presmoothing model was varied while treating the equating methods' assumptions about the unobservable test data as if they were true.

- This study evaluates the effect of presmoothing on the estimation of equated score differences and the standard errors of equated score differences (SEEDs) between chained and post-stratification equating functions. These evaluations, not considered in previous presmoothing and equating studies, are of interest because assessments of chained and post-stratification differences and their variability are important considerations for informing the selection of the chained or post-stratification equating function.

Method

This simulation study used a population data set and known population equating functions for equating test *X* to a more difficult test *Y* through an external anchor (*A*), where *X* and *A* were taken by population *P*, and *Y* and *A* were taken by a less able examinee population, *Q*. Four degrees of presmoothing and three sample sizes were considered. The biases and empirical variabilities of the chained kernel, chained equipercentile, post-stratification kernel and post-stratification equipercentile equating methods were evaluated with respect to their population equating functions. (A single criterion equating function was not used in this study.) The accuracies of the SEE and SEED estimates for the four equating methods and their differences were also evaluated. The four equating functions' biases and their SEE estimates were computed from the 4 x 3 = 12 presmoothing and sample size combinations based on 500 random samples from the population data.

Sample sizes. The *P* and *Q* data sets were each generated with equal sample sizes of 100, 200, and 1,000. Each sample size combination ($N_P = N_Q = 100$, $N_P = N_Q = 200$, and $N_P = N_Q = 1{,}000$) was replicated 500 times.

Degree of presmoothing. The frequencies from the data sets were used in their raw form and also presmoothed with three different loglinear presmoothing models (Holland & Thayer, 1987, 2000). The simplest loglinear model (M221, with 2 + 2 + 1 parameters) preserved the means and variances of the tests' and anchors' univariate distributions and the correlation between the test and anchor. Another loglinear model preserved the first six moments of the tests' and anchors' univariate distributions, along with the correlation between the test and anchor (M661, with 6 + 6 + 1 parameters). The third model was the actual population model fitted to the samples of data (MP). This 22-parameter model preserved the first four moments of the tests' and anchors' distributions, lumps at scores of zero for the tests and anchors, the frequencies and first three moments of the teeth scores (frequencies that were systematically lower than the overall distributions at every fifth score due to the use of the rounded formula scoring method), and four bivariate moments between the tests and anchors (von Davier et al., 2004, pp.159–167).

Population distributions. The population distributions were based on a smoothed version of test data that came from the large-volume administration of a verbal assessment (von Davier et al., 2004). The descriptive statistics from these populations, based on 10,634 examinees in *P* and 11,321 examinees in *Q*, are given in Tables 1 and 2. Figures 1–4 plot the marginal distributions from the two 22-parameter smoothing models. The modeled frequencies from the 22-parameter models fit the original raw data very well, with likelihood ratio χ^2 statistics of 1,966.9 (population *P*) and 1,896.0 (population *Q*) on 2,821 degrees of freedom.

Practical Issues

The sample generation from the *P* and *Q* population distributions used the following process. First bivariate cumulative probabilities were computed for each *XA* and *YA* score combination. Then a desired number of random uniform (0, 1) numbers were generated for the *P* and *Q* samples. For each of these random numbers, the *XA* and *YA* score combinations with the largest cumulative probabilities that were smaller than the random numbers were assigned. The resulting distributions were generally reflective of the populations' characteristics, but with random noise.

Table 1

Summary Statistics for XA in Population P

	X	A
Mean	39.25	17.05
SD	17.23	8.33
Skewness	– .11	– .01
Kurtosis	2.23	2.15
Min.	0	0
Max.	78	35

Note. Correlation = .88.

Table 2

Summary Statistics for YA in Population Q

	Y	A
Mean	32.69	14.39
SD	16.73	8.21
Skewness	.24	.26
Kurtosis	2.31	2.25
Min.	0	0
Max.	78	35

Note. Correlation = .87.

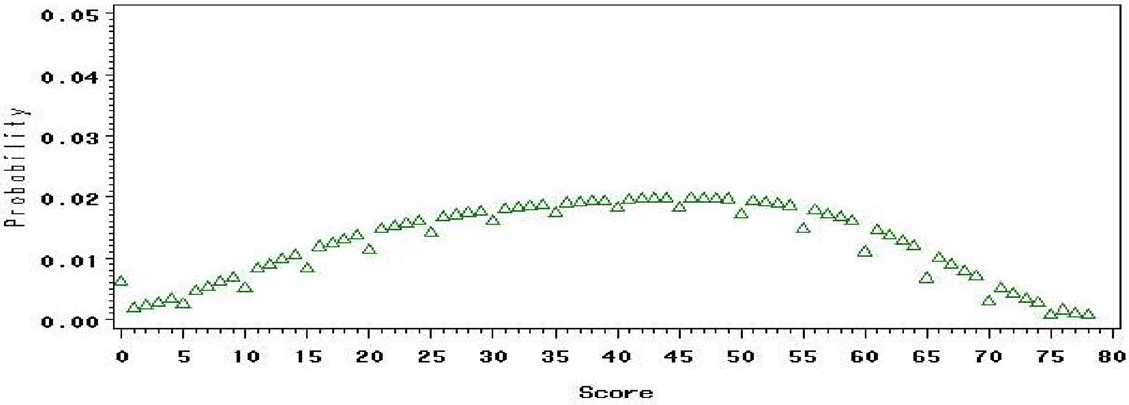

Figure 1. X_P**'s true marginal probability distribution.**

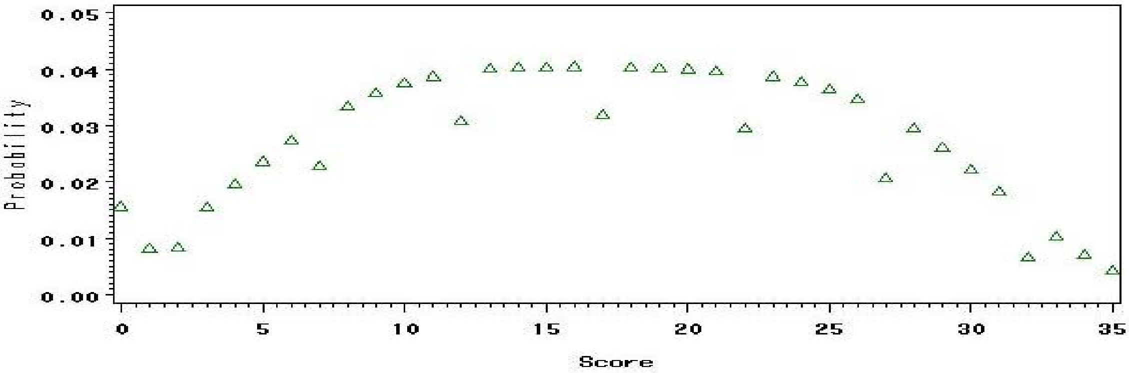

Figure 2. A_P**'s true marginal probability distribution.**

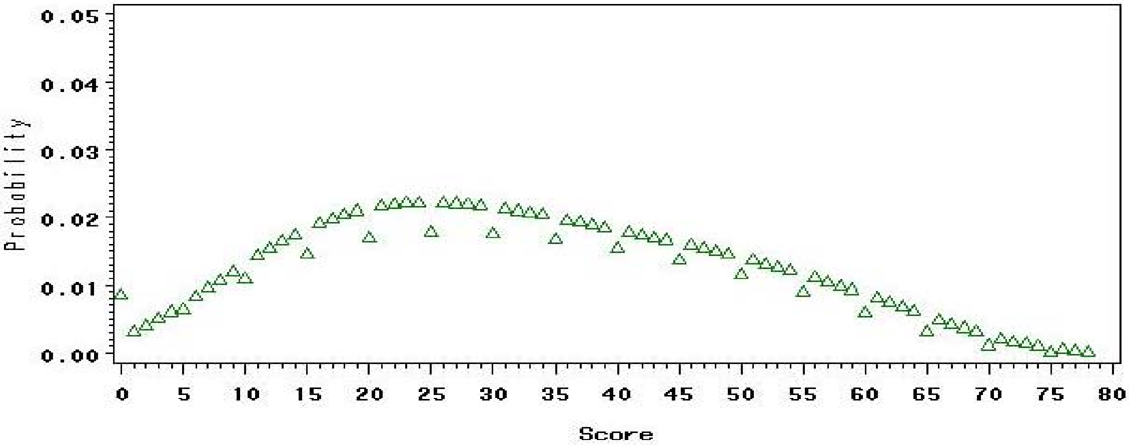

Figure 3. Y_Q**'s true marginal probability distribution.**

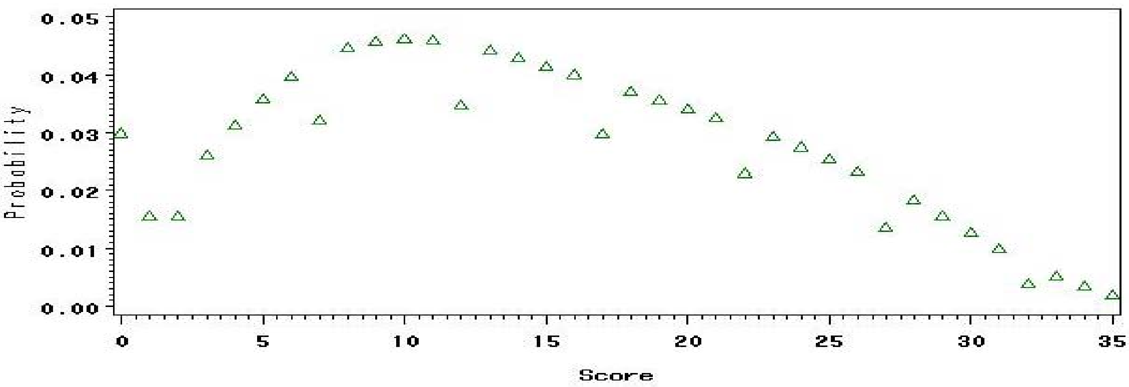

Figure 4. ***A_Q’s true marginal probability distribution.***

Post-stratification weights. Post-stratification equating is based on the direct equating of test *X* to test *Y* in a targeted synthetic population that is a mixture of populations *P* and *Q*, $T = wP + (1 – w)Q$. All post-stratification analyses in this study were conducted by setting $w = .5$.

Loglinear presmoothing. The 22-parameter models did not always converge in the study. To increase the convergence rates, the loglinear models used orthogonal polynomial score functions rather than the power functions that are typically described (Holland & Thayer, 1987, 2000). The use of orthogonal polynomials resulted in high convergence rates. The loglinear models converged for all 500 samples of size 1,000. For the samples of size 200, 9 out of the 500 did not converge. For the samples of size 100, 32 out of the 500 did not converge. The results for the samples of 200 are therefore based on the 491 converging *P* and *Q* models. The results for the samples of 100 are based on the 468 converging *P* and *Q* models.

Bandwidth selection. For the four observed distributions used in chained kernel equating and the two synthetic distributions used in post-stratification kernel equating, bandwidths were needed for estimating the Gaussian continuized cumulative density functions. These bandwidths were selected for each replication, degree of presmoothing and sample size. The bandwidth selection rule minimized the sum of two penalty functions (von Davier et al., 2004, pp. 61–64). The first penalty function was the sum of the squared differences between the continuized and discrete score probabilities. The second penalty function was used to minimize the number of scores (x_j) where the continuized distribution was U-shaped from x_j - 1/2 to x_j + 1/2. Bandwidths that minimized the sum of these two penalty functions resulted in density functions that were close reflections of the discrete distributions with very few modes. A parabolic interpolation procedure known as Brent’s method (Press, Teukolsky, Vetterling, & Flannery, 1992) was used to find the bandwidth that minimized the sum of the two penalty functions. When this procedure was used in the six population distributions, the selected bandwidths were 2.014 for X_P, 2.131 for Y_Q, 2.010 for A_P, 1.356 for A_Q, 1.895 for X_{P+Q}, and 1.977 for Y_{P+Q}. To provide meaningful bases for interpreting the values of the selected bandwidths, note that the bandwidth selected by minimizing the two penalty functions for an extremely smooth distribution tends to be approximately .5 or .6 (von Davier et al., 2004, pp. 106 & 124), while the bandwidth of a continuized distribution that retains only the mean and variance of the discrete distribution needs to be large (i.e., greater than 10 times the standard deviation of the distribution).

Analysis Strategy

The accuracies of the equating functions were evaluated as biases and variabilities from criterion equating functions. For bias, the average equated scores from the four equating methods (chained and post-stratification for both kernel and traditional equipercentile equating) for a particular combination of presmoothing and sample size were compared to the equated scores from that method using the population distributions. Weighted averages of the differences between average and population-equated scores were computed for the population probabilities of X_P. For example, the measure of overall bias in the chained equipercentile method using the M661 loglinear model and sample size $N_P = N_Q = 200$ was computed as:

$$\sum_j \left[\left(\sum_{iteration=1}^{491} \frac{e_{Y(ECE),M661,N200,iteration}(x_j)}{491} \right) - e_{Y(ECE),M22P,population}(x_j) \right] P(X_{P,M22P,population} = x_j) \quad (1)$$

(Note that the sample equating functions for sample sizes of $N_P = N_Q = 200$ were averaged over the 491 data sets where the loglinear models converged.)

Empirical variabilities were computed as weighted averages of the standard deviations of an equating method’s and continuization method’s equated scores for a particular combination of presmoothing and sample size across the population probabilities of X_P:

$$\sum_j \left[\sigma\left(e_{Y(ECE),M661,N200}(x_j) \right) \right] P(X_{P,M22P,population} = x_j) \quad . \qquad (2)$$

The delta method SEEs and SEEDs for each equating method, presmoothing, and sample size combination were also assessed for accuracy. The criterion for the SEEs and SEEDs were the standard deviations of all of the equated scores and equated score differences in the same conditions of equating method, presmoothing, and sample size:

$$\sum_{j}\left[\left(\sum_{iteration=1}^{491}\frac{SEE_{Y(ECE),M661,N200,iteration}(x_j)}{491}\right)-\sigma\left(e_{Y(ECE),M661,N200}(x_j)\right)\right]P(X_{P,M22P,population}=x_j). \quad (3)$$

Results

Evaluating the Smoothing

Table 3 gives the statistics of the population model's (MP) likelihood ratio chi-square statistics for each sample size of the *P* and *Q* distributions. The first row gives the statistics for 500 sums of 2,821 random chi-square deviates (what the likelihood ratio chi-square statistics estimate when the population MP model is fit to the sample data). The fit statistics from the actual data sets are much smaller than the simulated chi-square statistics and smallest when based on sample sizes of 100. The variability and ranges of the likelihood ratio chi-square statistics are also smaller than those of the simulated chi-square statistics.

Evaluating the Selected Bandwidths

Tables 4–7 give the descriptive statistics for the selected kernel bandwidths for the six distributions, the four presmoothing conditions, and sample size combination $N_P = N_Q = 1{,}000$. In these four tables, the columns labeled *Population value (MP)* show the bandwidths selected in the population data based on the MP loglinear model. The bandwidths selected from data that were presmoothed with the M221 (Table 4) and M661 (Table 5) models are much smaller than the population values and also smaller than the bandwidths selected from data presmoothed with the MP model (Table 6) and the bandwidths selected in the raw data (Table 7). The small size of the M221 and M661 selected bandwidths is due to the strongly presmoothed marginal distributions; only very local kernel continuizations were needed to produce sufficiently smooth continuized distributions that were close reflections of the discrete distributions.

Table 3

Likelihood Ratio Chi-Square Statistics for the Fits of the True, MP Models

	N	Replications	Mean	Median	SD	Skew	Min.	Max.
Simulated		500	2,812.239	2,813.547	77.698	0.069	2,597.163	3,022.066
XPAP	1,000	500	1,306.326	1,304.823	44.035	0.096	1,147.037	1,449.872
YQAQ	1,000	500	1,281.748	1,279.675	43.800	0.022	1,126.264	1,439.312
XPAP	200	491	696.691	696.861	23.223	0.056	634.903	774.560
YQAQ	200	491	687.334	688.273	22.958	–0.192	618.860	749.749
XPAP	100	468	460.337	460.706	17.076	–0.053	407.190	509.899
YQAQ	100	468	456.520	456.864	16.269	–0.106	399.064	505.585

Note. *Df* = 2,821.

Table 4

Bandwidth Statistics, M221

Distribution	Population value (MP)	*n*	Mean	Median	SD	Skew	Min.	Max.
X_P	2.014	500	0.608	0.608	0.006	0.117	0.592	0.629
Y_Q	2.131	500	0.594	0.593	0.006	0.100	0.575	0.613
A_P	2.010	500	0.563	0.563	0.006	-0.104	0.547	0.580
A_Q	1.356	500	0.542	0.542	0.005	-0.199	0.521	0.556
X_{P+Q}	1.895	500	0.606	0.606	0.005	0.079	0.590	0.626
Y_{P+Q}	1.977	500	0.601	0.601	0.005	0.172	0.584	0.620

Note. *N* = 1,000.

Table 5

Bandwidth Statistics, M661

Distribution	Population value (MP)	*n*	Mean	Median	SD	Skew	Min.	Max.
X_P	2.014	500	1.299	0.165	1.455	2.181	0.576	8.158
Y_Q	2.131	500	0.629	0.591	0.160	6.237	0.557	2.570
A_P	2.010	500	1.411	0.916	1.002	0.842	0.533	4.319
A_Q	1.356	500	0.703	0.544	0.232	1.281	0.512	1.853
X_{P+Q}	1.895	500	1.137	0.605	1.220	2.559	0.561	6.897
Y_{P+Q}	1.977	500	0.783	0.601	0.698	4.577	0.569	5.735

Note. N = 1,000.

Table 6

Bandwidth Statistics, MP

Distribution	Population value (MP)	*n*	Mean	Median	SD	Skew	Min.	Max.
X_P	2.014	500	1.907	1.914	0.308	0.532	0.868	4.349
Y_Q	2.131	500	1.849	1.871	0238	-0.564	0.958	2.377
A_P	2.010	500	1.849	1.824	0.350	1.243	0.952	3.745
A_Q	1.356	500	1.367	1.363	0.244	0.169	0.701	2.246
X_{P+Q}	1.895	500	1.987	2.013	0.298	-0.253	1.134	2.744
Y_{P+Q}	1.977	500	1.944	1.963	0.246	-0.334	1.054	2.625

Note. N = 1,000.

Table 7

Bandwidth Statistics, Raw

Distribution	Population value (MP)	*n*	Mean	Median	SD	Skew	Min.	Max.
X_P	2.014	500	4.084	3.925	1.052	0.845	2.009	7.788
Y_Q	2.131	500	3.580	3.529	0.808	0.695	1.850	6.775
A_P	2.010	500	2.275	2.151	0.547	0.802	1.172	4.122
A_Q	1.356	500	1.796	1.742	0.360	0.816	1.066	3.408
X_{P+Q}	1.895	500	4.262	4.090	1.114	0.654	2.229	7.519
Y_{P+Q}	1.977	500	4.081	4.020	0.917	0.605	2.300	7.264

Note. *N* = 1,000.

The bandwidths selected from the MP models (Table 6) are closer to the population values than those of the raw data and the M221 and M661 models. The bandwidths selected from the raw data (Table 7) are substantially larger than those from the population bandwidths, suggesting that the part of the bandwidth selection rule that limits the number of modes in the continuized distribution was actually smoothing out some of the roughness of the raw data.

Bias and Empirical Variability of the Equating Functions

Tables 8–11 give the bias and the empirical variability of the equating functions across sample sizes and presmoothing models. Figures 5–8 plot the equating functions' bias series across presmoothing models for the sample size combination $N_P = N_Q = 200$. The biases are small relative to variability and were as expected for the four smoothing conditions. The equating functions based on the raw data and on the data that were presmoothed based on the population MP model exhibited the smallest biases. The M221 model created the most bias across the equating functions and sample sizes out of all of the presmoothing conditions. From Figures 5–8, the M221 model can be seen to have created both over- and under-estimation problems with respect to all four of the population-equating functions. The raw data created some overestimation problems for the most extreme scores of the chained equipercentile function (Figure 5) but underestimation problems for the highest scores of the chained kernel function (Figure 7). The chained kernel's bias (Figure 7) in the high end of the score range shows underestimation of approximately one score point for all of the presmoothed and raw data conditions, which is slightly greater and more consistent tail-bias than was present with the other equating methods. The bias series for the traditional equipercentile functions exhibited small and abrupt shifts corresponding to the teeth structures in the original data (Figures 5–6), while those for the kernel functions did not (Figures 7–8).

In terms of empirical variability, the equating functions were most variable when based on the raw data, followed by M661, then by MP, and then by M221 (Tables 8–11). The high variability of the M661 equating functions relative to the MP equating functions is interesting. M661 is simpler than MP by 22 – 13 = 9 total parameters but is actually over-parameterized by two moments in each of the four marginal distributions. These fifth and sixth moments did not exist in the population distributions, and preserving them in the sample data sets apparently resulted in more noise being added to the equating functions. Comparing equating functions' empirical variabilities, the kernel equating functions are less variable than the traditional equipercentile functions. The kernel and equipercentile post-stratification equating functions are less variable than the chained equating functions.

Table 8

Bias and Empirical Variability for Chained Equipercentile

Bias	$N_P = N_Q =$			Empirical variability	$N_P = N_Q =$		
	1,000	200	100		1,000	200	100
M221	–0.035	–0.058	–0.001	M221	0.495	1.182	1.639
M661	–0.005	–0.030	0.039	M661	0.722	1.709	2.382
MP	–0.006	–0.031	0.027	MP	0.660	1.574	2.162
Raw	–0.003	0.010	0.159	Raw	0.905	2.086	2.842

Note. All $X_{P.}$

Table 9

Bias and Empirical Variability for Post-Stratification Equipercentile

Bias	$N_P = N_Q =$			Empirical variability	$N_P = N_Q =$		
	1,000	200	100		1,000	200	100
M221	–0.021	–0.043	0.020	M221	0.479	1.143	1.578
M661	–0.006	–0.032	0.036	M661	0.646	1.540	2.159
MP	–0.006	–0.033	0.028	MP	0.599	1.435	1.997
Raw	–0.009	–0.021	0.018	Raw	0.778	1.820	2.599

Note. All $X_{P.}$

Table 10

Bias and Empirical Variability for Chained Kernel

Bias	$N_P = N_Q =$			Empirical variability	$N_P = N_Q =$		
	1,000	200	100		1,000	200	100
M221	–0.037	–0.060	–0.003	M221	0.496	1.183	1.640
M661	–0.005	–0.031	0.021	M661	0.718	1.629	2.159
MP	–0.007	–0.032	0.022	MP	0.633	1.484	1.992
Raw	–0.005	–0.028	0.033	Raw	0.696	1.506	1.994

Note. All $X_{P.}$

Table 11

Bias and Empirical Variability for Post-Stratification Kernel

Bias	$N_P = N_Q =$			Empirical	$N_P = N_Q =$		
	1,000	200	100	variability	1,000	200	100
M221	–0.024	–0.045	0.017	M221	0.479	1.143	1.578
M661	–0.008	–0.033	0.029	M661	0.636	1.463	1.950
MP	–0.007	–0.035	0.027	MP	0.588	1.385	1.902
Raw	–0.007	–0.040	–0.009	Raw	0.623	1.412	2.027

Note. All $X_{P.}$

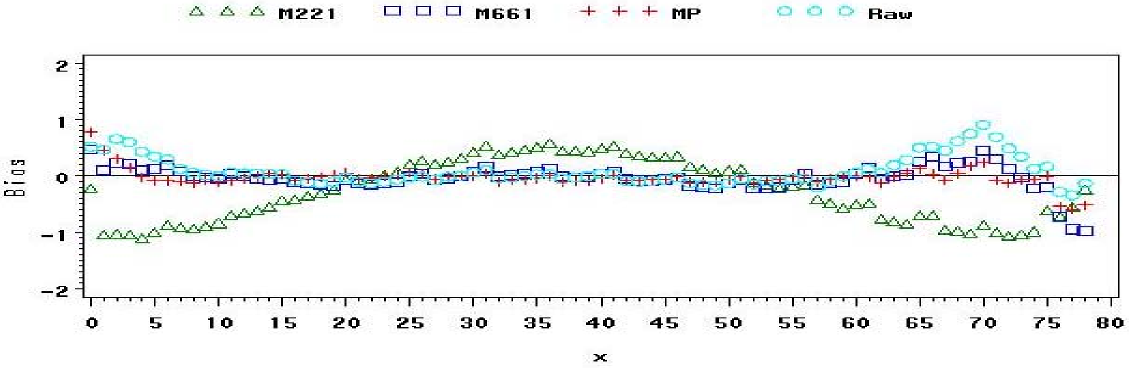

Figure 5. **Chained equipercentile bias.**

Note. *Np* = *Nq* = 200.

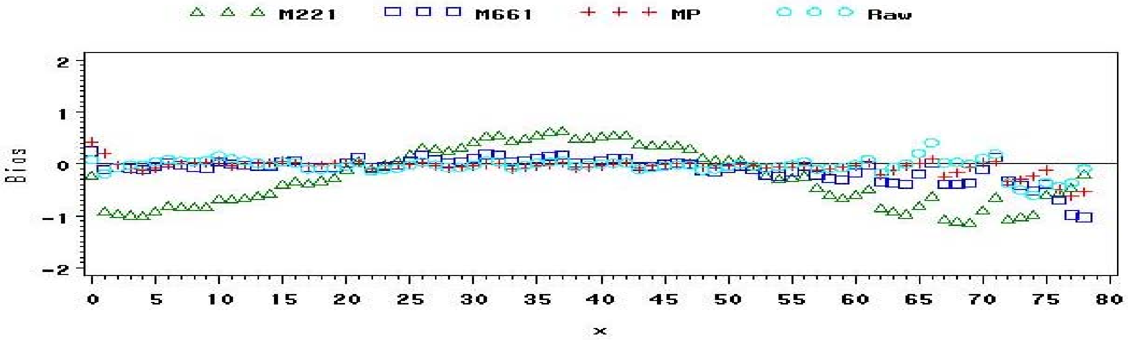

Figure 6. **Post-stratification equipercentile bias.**

Note. *Np* = *Nq* = 200.

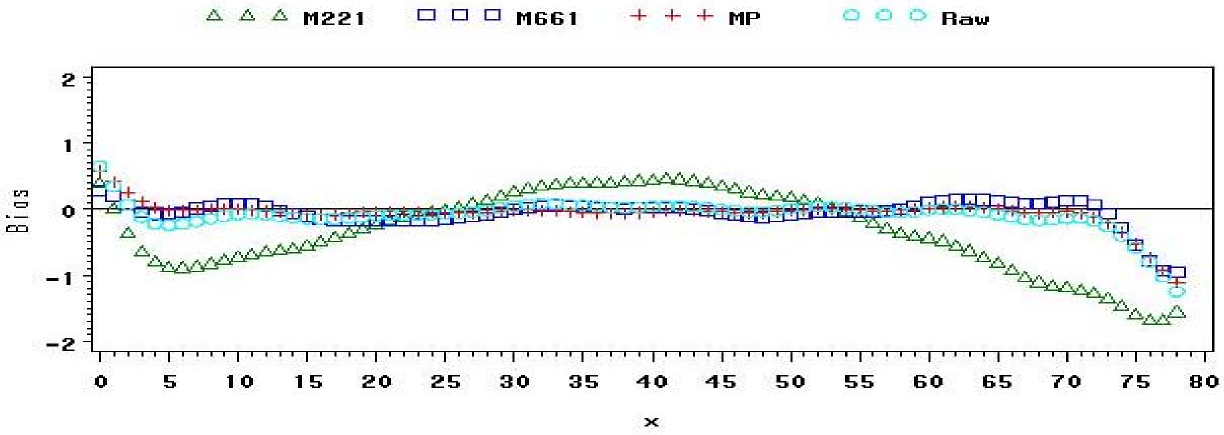

Figure 7. **Chained kernel bias.**

Note. *Np* = *Nq* = 200.

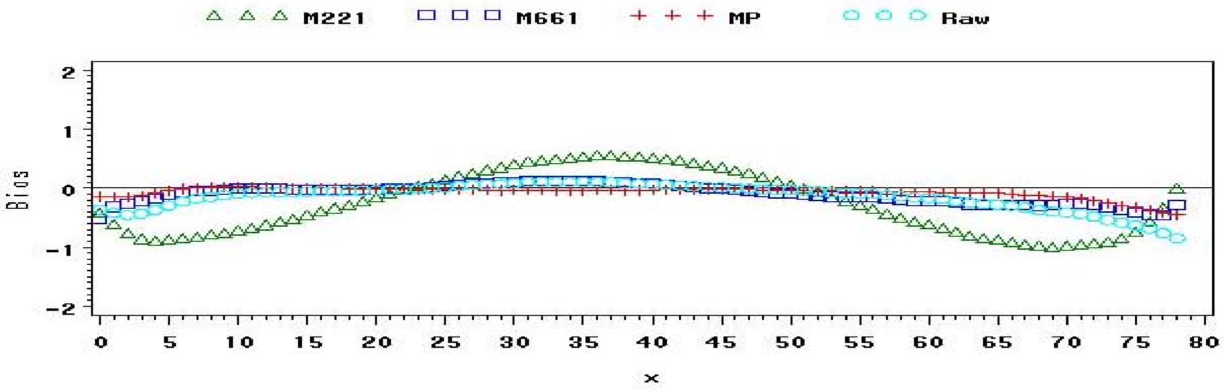

Figure 8. **Post-stratification kernel bias.**

Note. *Np* = *Nq* = 200.

Accuracy of the SEEs

Tables 12–15 describe how well the SEEs estimated the actual variability in the equating functions across sample sizes and presmoothing models. Figures 9–24 plot the average SEEs and the actual equating function standard deviations across presmoothing models for sample size combinations $N_P = N_Q = 200$ and $N_P = N_Q = 1{,}000$. For samples of 100 and 200, the post-stratification SEEs consistently underestimated actual variability (Tables 13 and 15). The M661 model created substantial estimation problems for the chained kernel method (Table 14). The source of M661’s estimation problems was the sparseness of data on Y_Q at the high end of the score range, which resulted in divisions by very small numbers in the standard error formulas and very large SEEs for the highest X_P scores (Figure 18).

The average MP SEEs at the X_P score of 0 (4.09) for the chained equipercentile method overestimated actual variability (1.87) for sample sizes of 200 (Figure 11). The source of this overestimation was nine extremely large (> 30) SEEs. Further inspections of the raw data of these particular replications showed that at least two of the four marginal distributions had no observations at scores of zero, where the lumps that truly existed in the populations were incorporated in the MP models and the extremely small smoothed values were then used as denominators for the standard error formulas. When the average SEE at $X_P = 0$ was computed without the extreme SEEs of these nine replications, it was much closer to the actual standard deviation of the corresponding chained equating function (1.74 vs. 1.87).

Table 12

Accuracy of the Delta Method Standard Errors for Chained Equipercentile

SEEs	$N_P = N_Q =$		
	1,000	200	100
M221	0.024	–0.025	–0.006
M661	0.024	–0.047	–0.035
MP	0.010	–0.066	0.058
Raw	0.012	–0.006	0.188

Note. All $X_{P.}$

Table 13

Accuracy of the Delta Method Standard Errors for Post-Stratification Equipercentile

SEEs	$N_P = N_Q =$		
	1,000	200	100
M221	0.013	–0.047	–0.031
M661	0.012	–0.077	–0.099
MP	0.006	–0.084	–0.083
Raw	–0.009	–0.205	–0.618

Note. All $X_{P.}$

Table 14

Accuracy of the Delta Method Standard Errors for Chained Kernel

SEEs	$N_P = N_Q =$		
	1,000	200	100
M221	0.024	–0.026	–0.006
M661	8.29 X 10^{100}	7.56 X 10^{105}	1.61 X 10^{99}
MP	0.007	–0.072	–0.034
Raw	0.002	–0.070	–0.052

Note. All $X_{P.}$

Table 15

Accuracy of the Delta Method Standard Errors for Post-Stratification Kernel

SEEs	$N_P = N_Q =$		
	1,000	200	100
M221	0.013	–0.047	–0.031
M661	0.017	–0.046	–0.015
MP	0.007	–0.071	–0.064
Raw	–0.004	–0.197	–0.591

Note. All $X_{P.}$

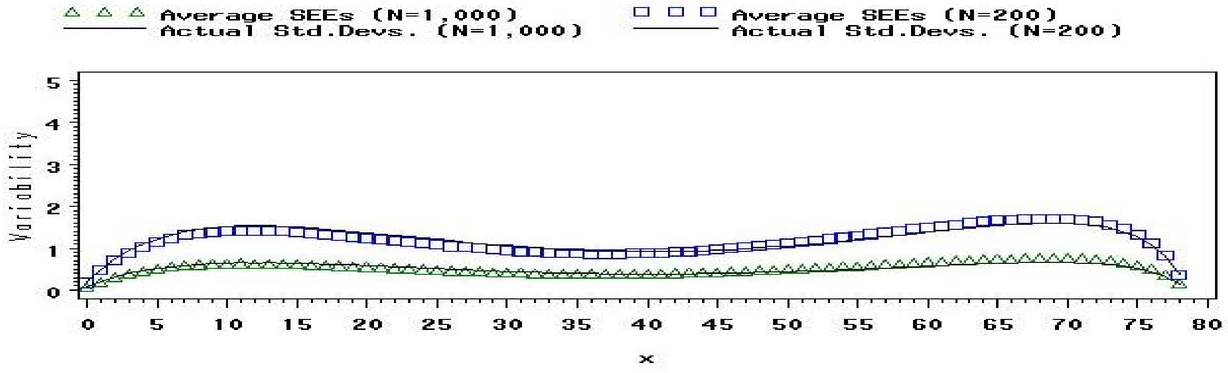

Figure 9. **Chained equipercentile SEE, M221.**

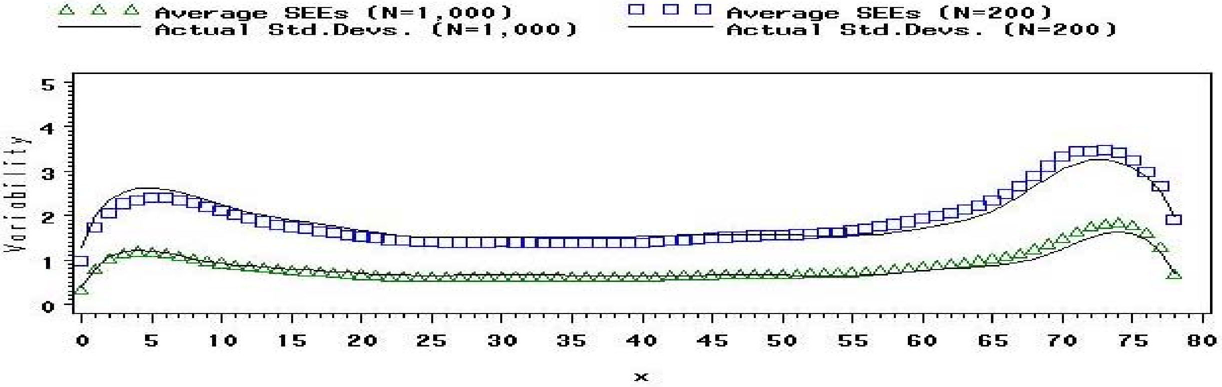

Figure 10. **Chained equipercentile SEE, M661.**

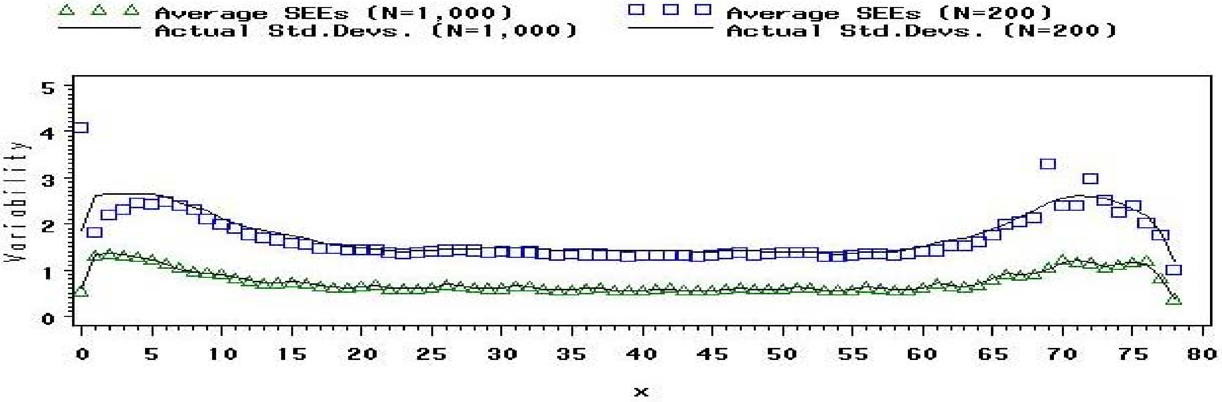

Figure 11. **Chained equipercentile SEE, MP.**

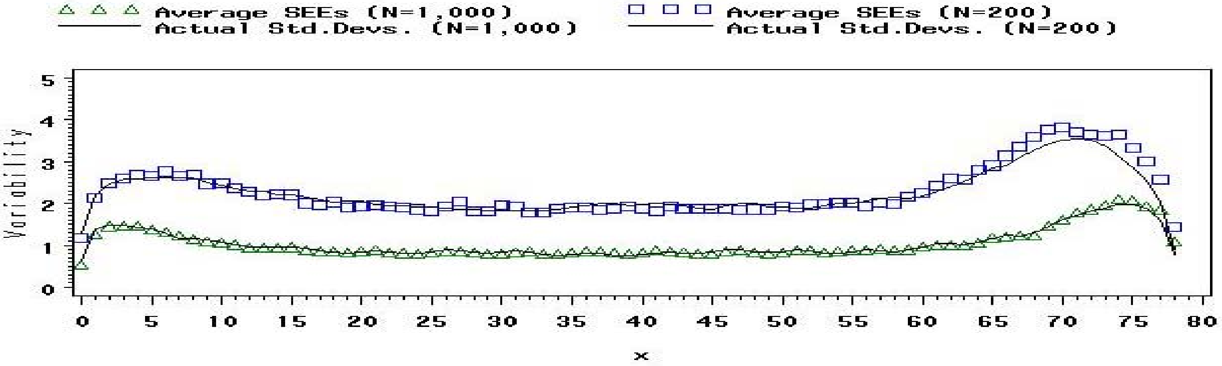

Figure 12. **Chained equipercentile SEE, raw.**

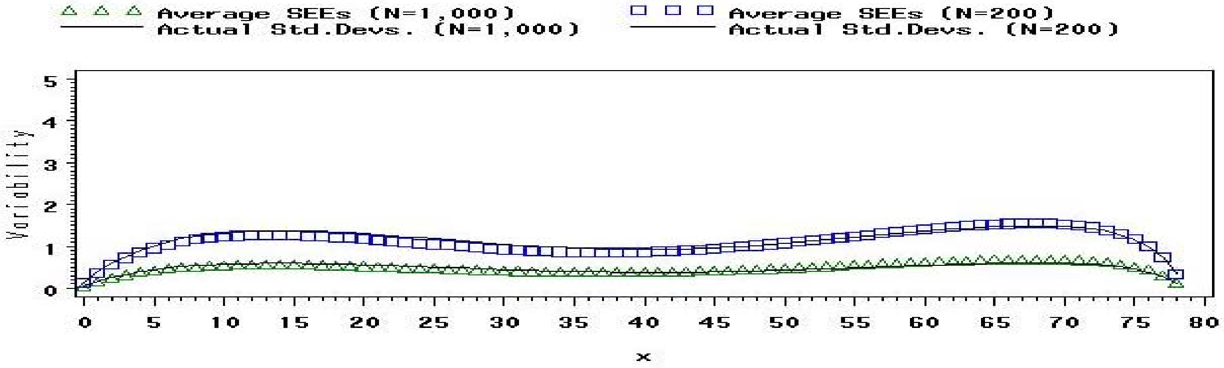

Figure 13. **Post-stratification equipercentile SEE, M221.**

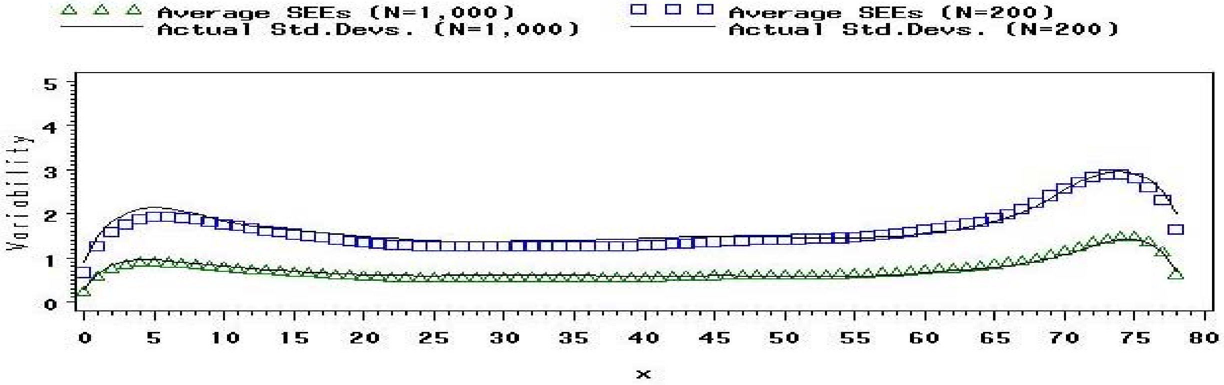

Figure 14. **Post-stratification equipercentile SEE, M661.**

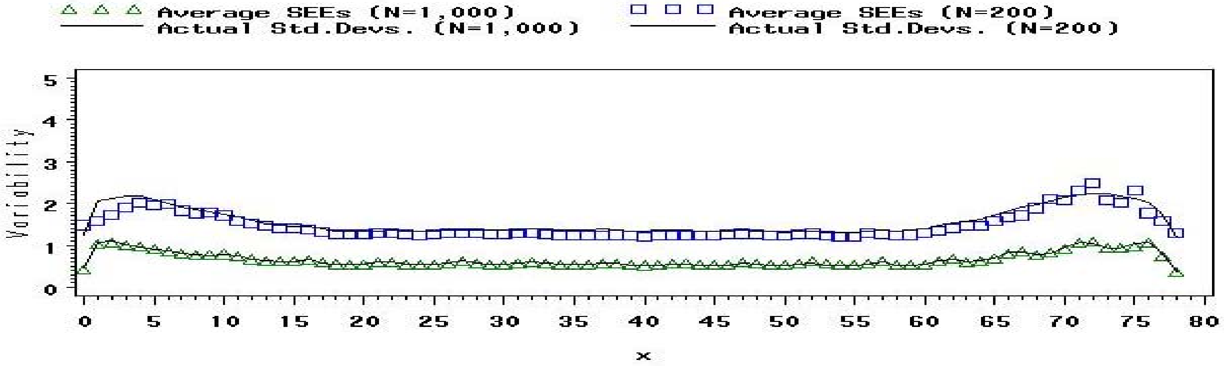

Figure 15. **Post-stratification equipercentile SEE, MP.**

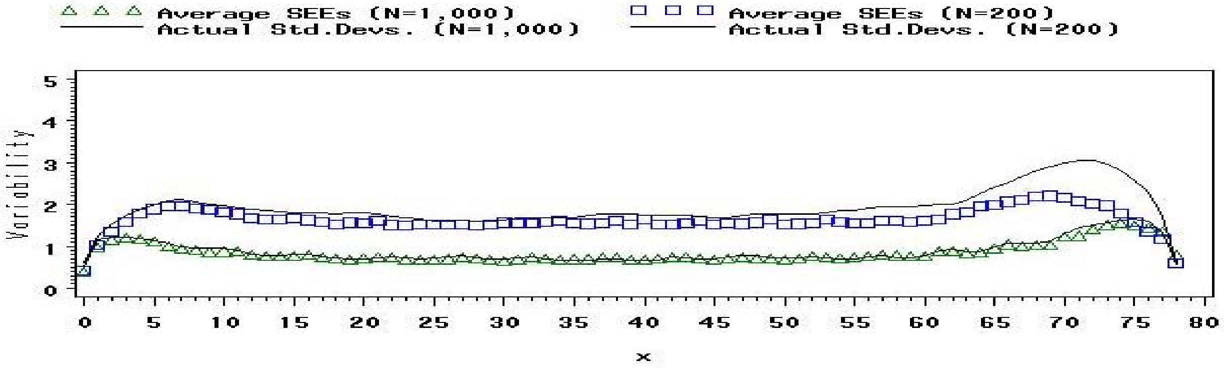

Figure 16. **Post-stratification equipercentile SEE, raw.**

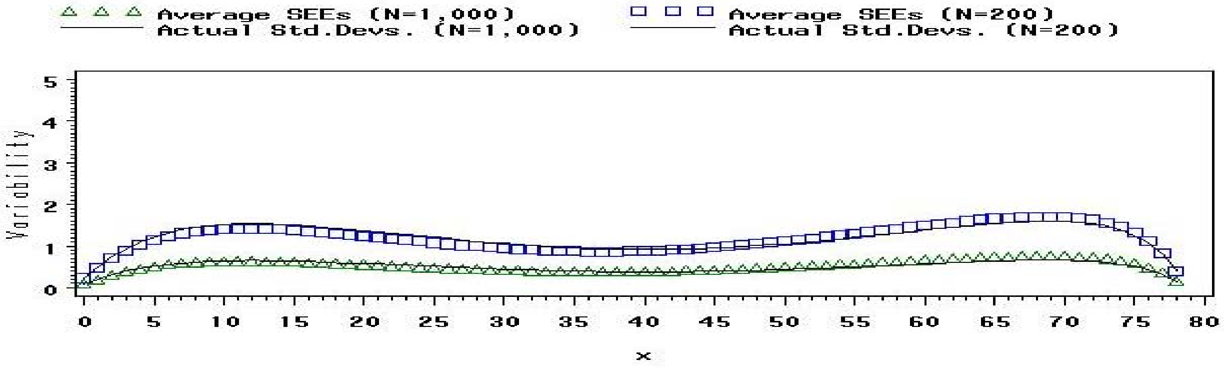

Figure 17. **Chained kernel SEE, M221.**

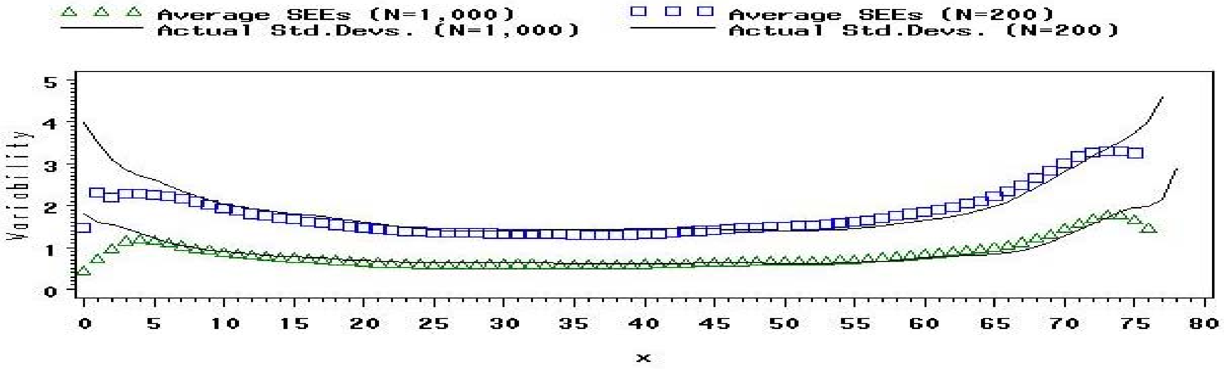

Figure 18. **Chained kernel SEE, M661.**

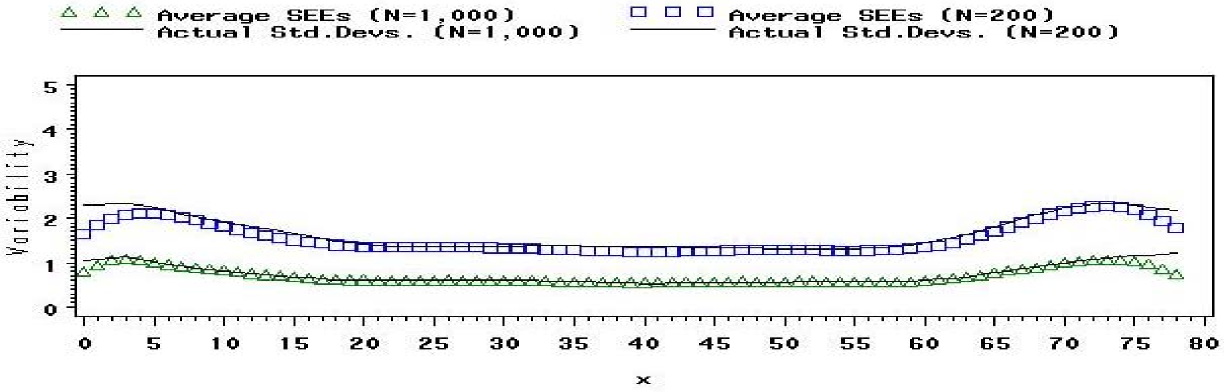

Figure 19. **Chained kernel SEE, MP.**

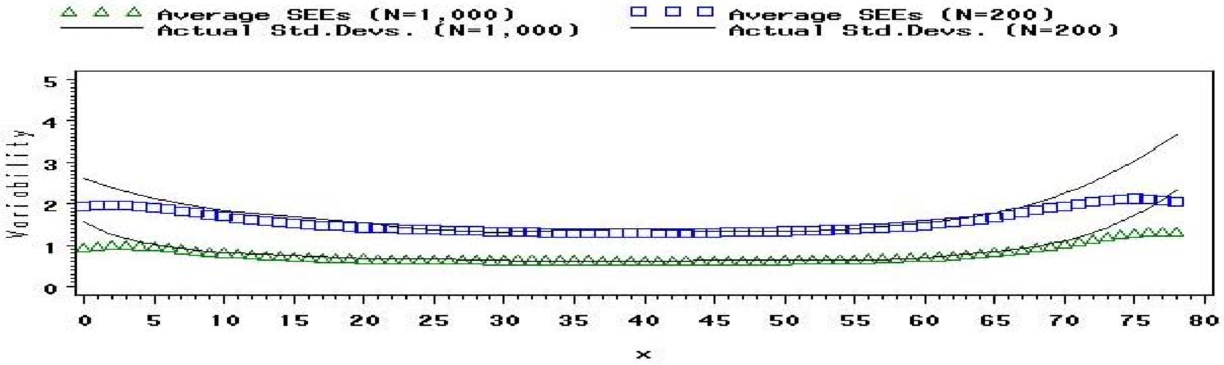

Figure 20. **Chained kernel SEE, raw.**

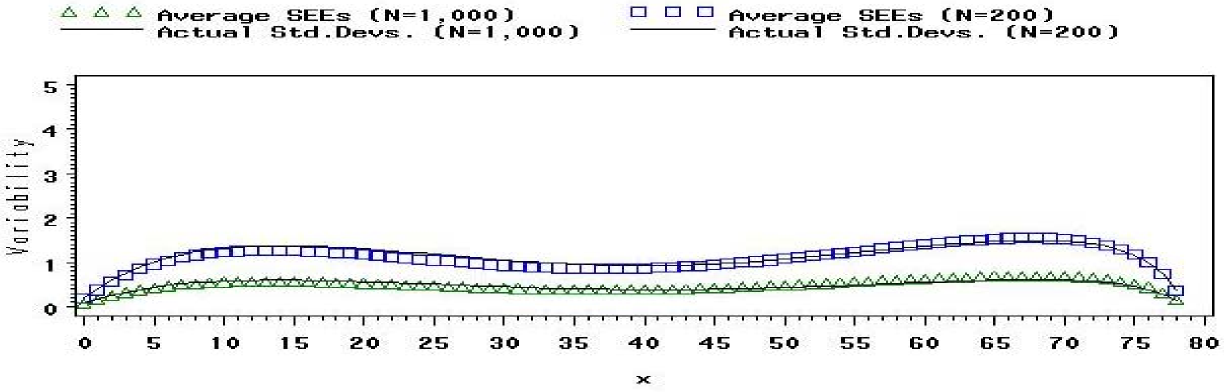

Figure 21. **Post-stratification kernel SEE, M221.**

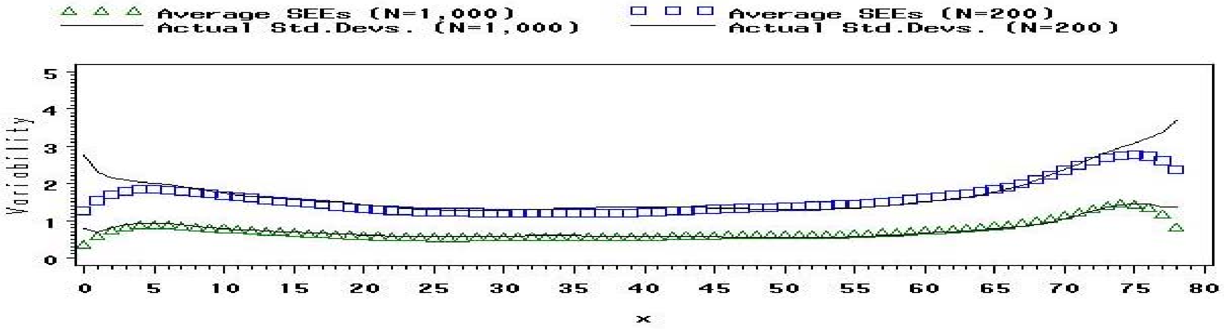

Figure 22. **Post-stratification kernel SEE, M661.**

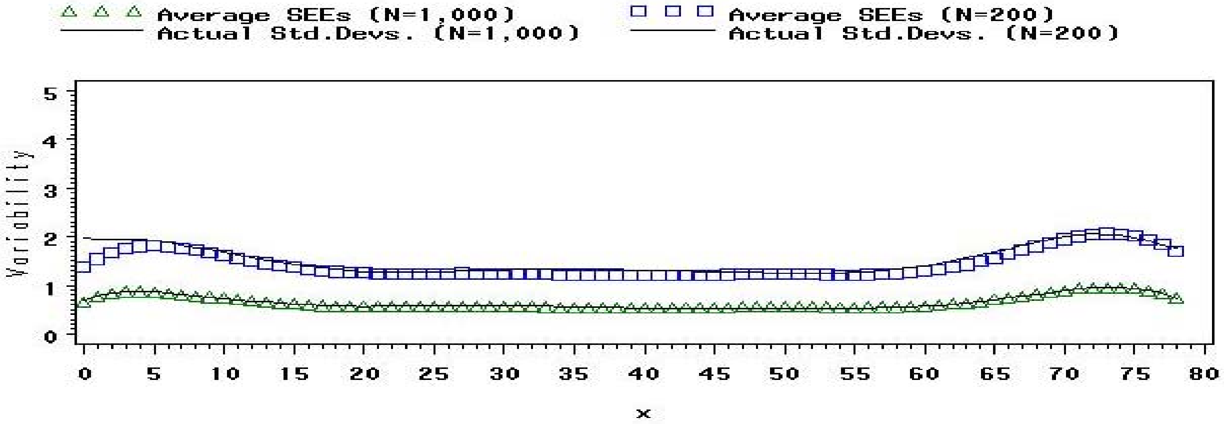

Figure 23. **Post-stratification kernel SEE, MP.**

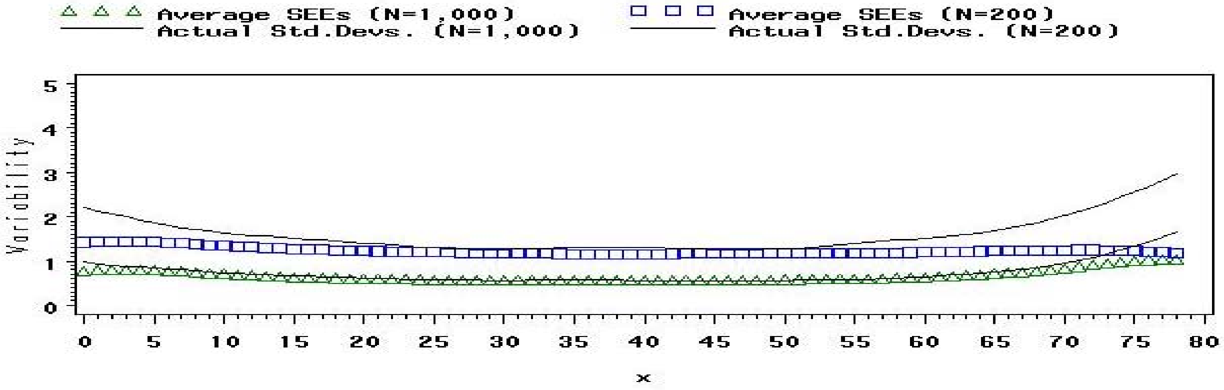

Figure 24. **Post-stratification kernel SEE, raw.**

Figures 9–24 show that SEEs (shown as squares and triangles) are generally close approximations to actual variability (shown as solid lines) for most of the score range. The less accurate SEEs are at the tails of the score ranges and at the smaller sample sizes. Similar to the series of bias figures (Figures 5–8), the SEE series for the equipercentile functions based on raw data and the MP model exhibit small and abrupt shifts corresponding to the teeth structures in the original data (Figures 11, 12, 15, and 16) while the raw and MP SEEs of the kernel functions do not (Figures 19, 20, 23, and 24). The SEEs based on any of the presmoothing models and sample sizes could be characterized as sufficiently accurate for practical use, so long as the variability estimates in the tails are not taken too seriously.

Bias and Empirical Variability of Equating Function Differences

The biases and empirical variabilities of the chained and post-stratification differences are summarized across all sample size combinations and presmoothing models in Tables 16–17. Figures 25 and 26 plot the bias series of the kernel and equipercentile chained and post-stratification differences across presmoothing models for the sample size combination $N_P = N_Q =$ 200. MP produced the smallest biases, and M221 produced the largest biases. In terms of empirical variability, the most to least variable differences were based on raw data, M661, MP, and finally M221. The variabilities of equating function differences were of smaller magnitude than the individual equating functions' variabilities.

Table 16

Bias and Variability for Equipercentile Chained and Post-Stratification Differences

Bias	$N_P = N_Q =$			Empirical	$N_P = N_Q =$		
	1,000	200	100	variability	1,000	200	100
M221	–0.014	–0.015	–0.021	M221	0.157	0.377	0.543
M661	0.001	0.002	0.003	M661	0.350	0.808	1.192
MP	0.000	0.001	–0.001	MP	0.298	0.694	1.014
Raw	0.006	0.031	0.141	Raw	0.535	1.390	2.201

Note. All $X_{P.}$

Table 17

Bias and Variability for Kernel Chained and Post-Stratification Differences

Bias	$N_P = N_Q =$			Empirical	$N_P = N_Q =$		
	1,000	200	100	variability	1,000	200	100
M221	–0.013	–0.014	–0.020	M221	0.157	0.378	0.543
M661	0.003	0.003	–0.008	M661	0.367	0.764	1.028
MP	0.000	0.003	–0.005	MP	0.248	0.568	0.814
Raw	0.002	0.012	0.042	Raw	0.354	0.782	1.188

Note. All X_P.

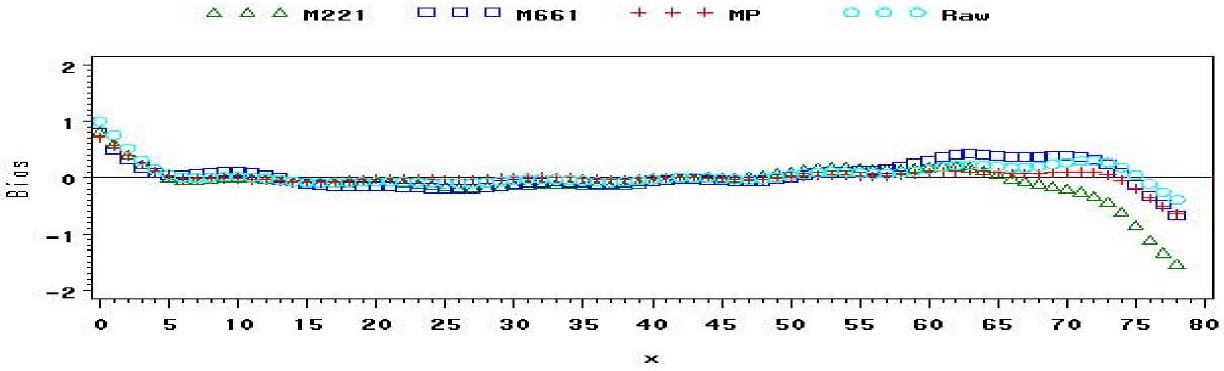

Figure 25. **Chained and post-stratification kernel bias.**

Note. *Np* = *Nq* = 200.

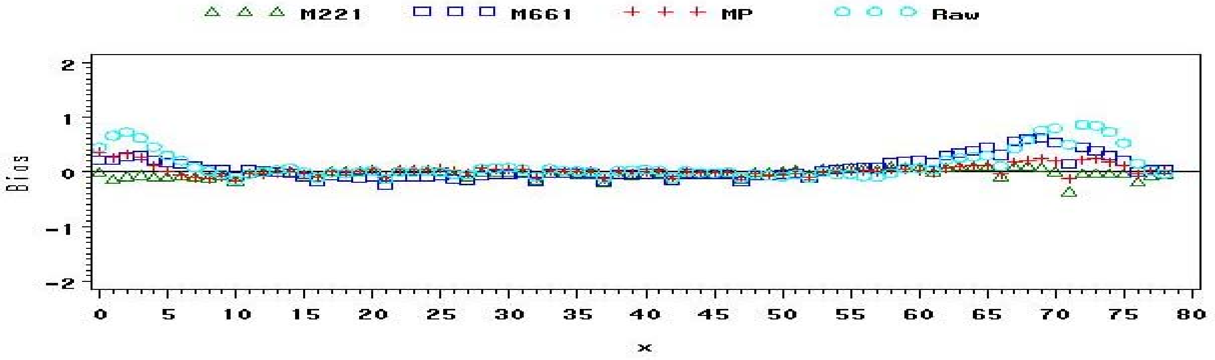

Figure 26. **Chained and post-stratification equipercentile bias.**

Note. *Np* = *Nq* = 200.

Accuracy of the SEEDs

The accuracies of the equipercentile and kernel chained and post-stratification SEEDs are summarized across all sample size combinations and presmoothing models in Tables 18–19. Figures 27–34 summarize the chained post-stratification SEED accuracy for the equipercentile and kernel functions for sample size combinations $N_P = N_Q = 200$ and $N_P = N_Q = 1{,}000$. Most of the features of these SEEDs have been described in the SEE descriptions presented earlier. The unique result is that the raw chained and post-stratification SEEDs that are based on equipercentile equating functions greatly overestimate the actual variabilities of the equating function differences (Figure 30). Some follow-up analyses that focused on approaches to continuization were performed to explain the relatively poor performance of the raw equipercentile SEEDs. These analyses are described in the discussion section.

Table 18

Accuracy of the Delta Method Standard Errors for Equipercentile Chained and Post-Stratification SEED

SEEs	$N_P = N_Q =$		
	1,000	200	100
M221	0.012	–0.001	–0.008
M661	0.001	–0.008	–0.030
MP	0.017	0.086	0.269
Raw	0.993	1.792	1.829

Note. All $X_{P.}$

Table 19

Accuracy of the Delta Method Standard Errors for Kernel Chained and Post-Stratification SEED

SEEs	$N_P = N_Q =$		
	1,000	200	100
M221	0.011	–0.002	–0.010
M661	8.3 X 10^{100}	7.6 X 10^{105}	1.6 X 10^{99}
MP	–0.005	–0.019	–0.028
Raw	–0.018	0.026	0.044

Note. All $X_{P.}$

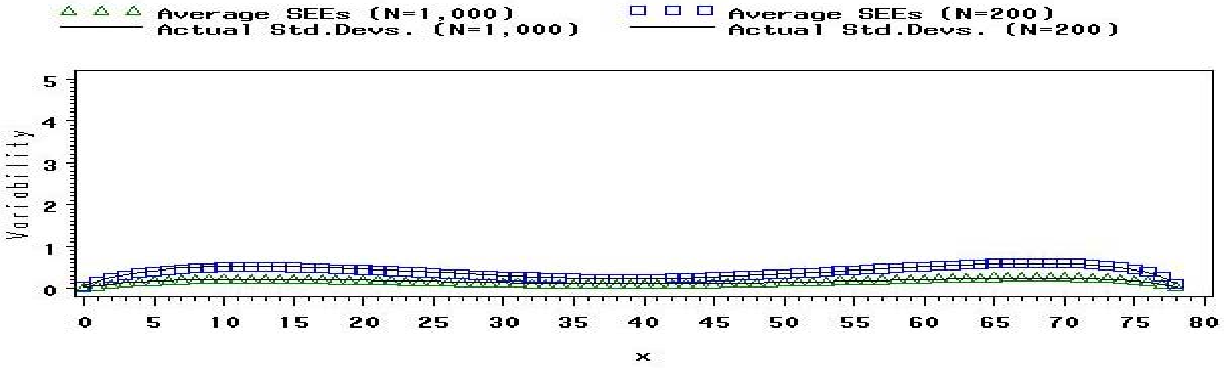

Figure 27. **Chained and post-stratification equipercentile SEED, M221.**

Note. *Np* = *Nq* = 200.

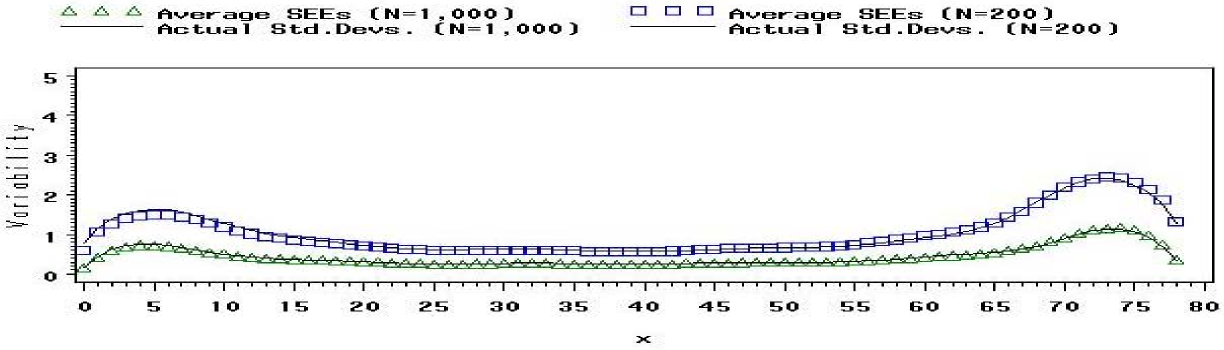

Figure 28. **Chained and post-stratification equipercentile SEED, M661.**

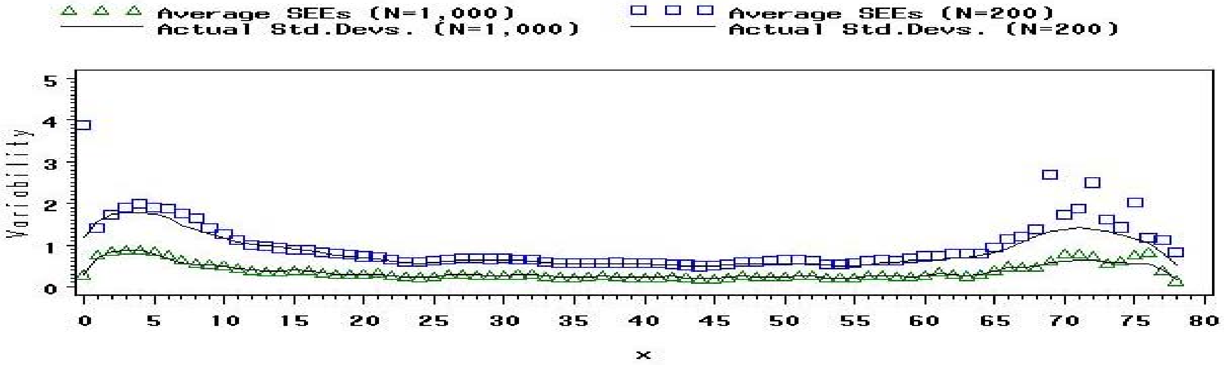

Figure 29. **Chained and post-stratification equipercentile SEED, MP.**

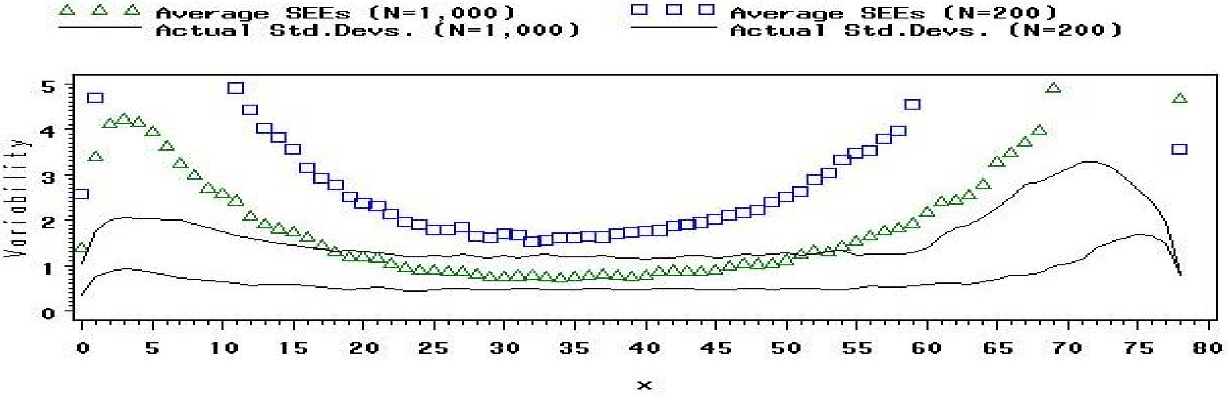

Figure 30. **Chained and post-stratification equipercentile SEED, raw.**

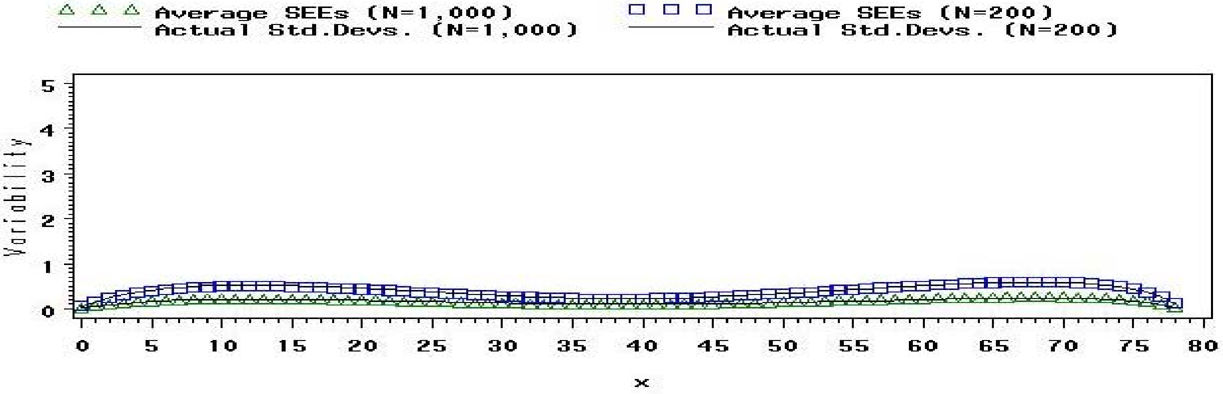

Figure 31. **Chained and post-stratification kernel SEED, M221.**

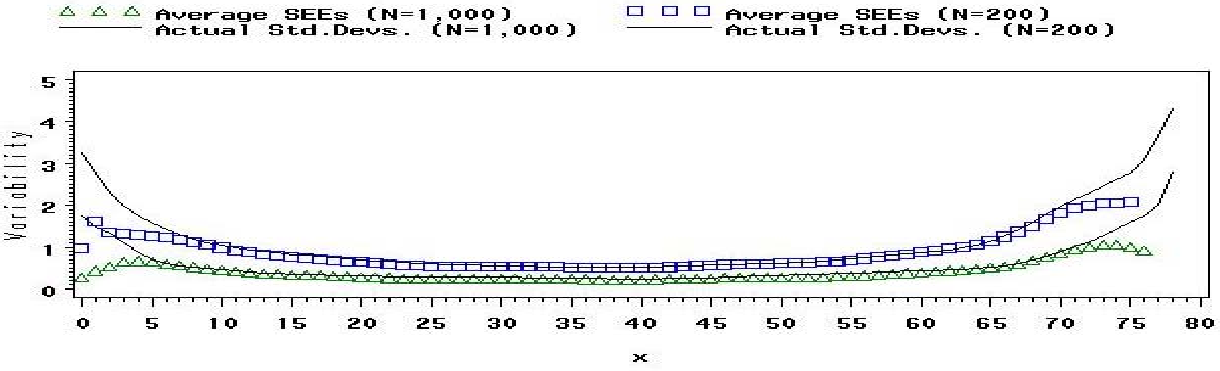

Figure 32. **Chained and post-stratification kernel SEED, M661.**

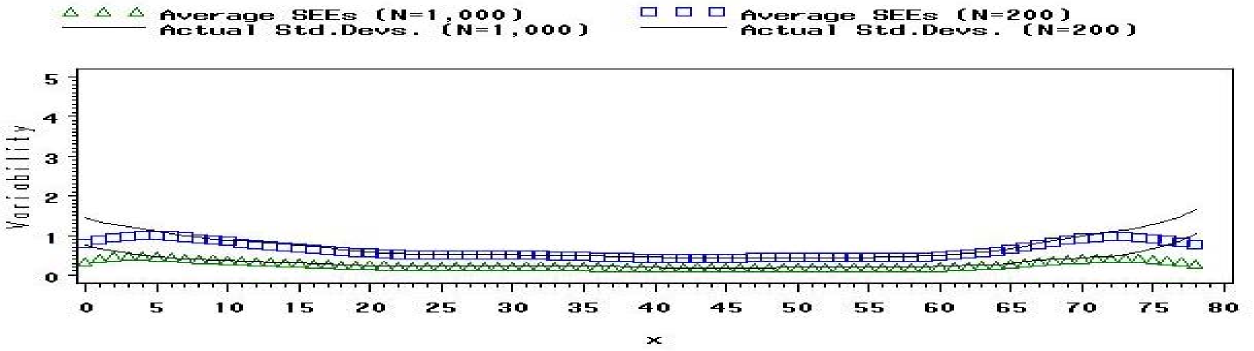

Figure 33. **Chained and post-stratification kernel SEED, MP.**

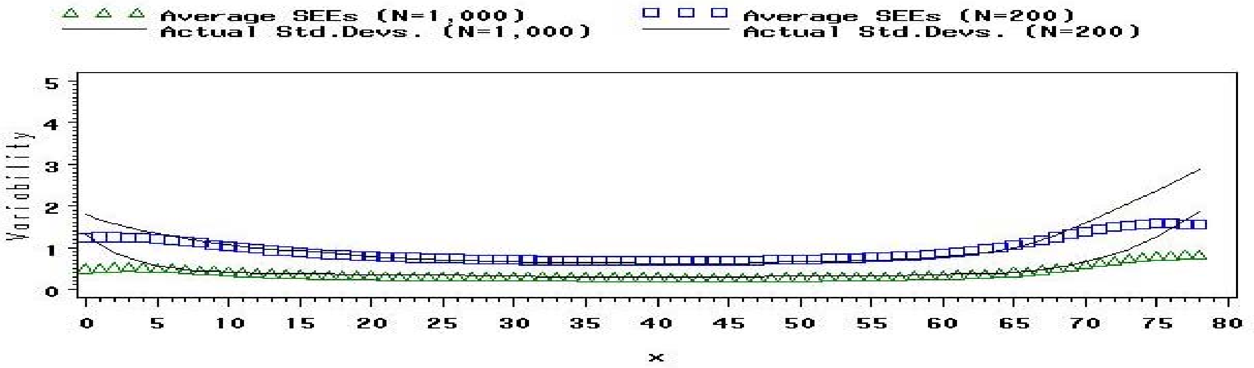

Figure 34. **Chained and post-stratification kernel SEED, raw.**

Discussion

The purpose of this study was to evaluate the impact of the correctness of the loglinear presmoothing model on equating function and SEE accuracy for NEAT equating. Prior studies of loglinear presmoothing have focused on either equating function accuracy (Hanson, 1991; Hanson et al., 1994; Livingston, 1993; Skaggs, 2004) or on SEE accuracy (Liou & Cheng, 1995; Liou et al., 1997), suggesting that the correctness of the presmoothing model is important for equating function accuracy but possibly not important for SEE accuracy. The evaluation in this study was across sample sizes, NEAT equating methods (chained and post-stratification), and continuization methods (percentile-rank and kernel). In addition, the accuracy of equating function differences and SEEDs for chained and post-stratification differences was also evaluated.

The results of this study on equating function accuracy replicated previous findings of a bias-variability tradeoff for given degrees of presmoothing. This tradeoff was most clearly seen in terms of two incorrect presmoothing models. All of the simulated data were generated from an extremely complex 22-parameter model. In terms of this population model, the M221 model was underparameterized, and the M661 model was overparameterized in the marginal distributions. The equating functions based on the M221 model were the least variable and most biased of all the equating functions. The equating functions based on the M661 model were less biased than those from the M221 model but more variable than those from the population MP model. Both the M221 and M661 models ignored the teeth and lumps at score zero and could be regarded as models that ‘ground down the teeth’ in the presmoothing step of equating (von Davier et al., 2004, p. 64). The bias and variability problems of the M221 and M661 models reiterate the von Davier et al. recommendation to get a correct smoothing model in the presmoothing step and then select a kernel bandwidth that eliminates the teeth in the continuization step. The use of raw data did not have any bias advantages over the two reasonable smoothing models (M661 and MP). The implications of less- and more- parameterized presmoothing models on equating function accuracy directly apply to the estimation accuracies of chained and post-stratification equating function differences.

The evaluation of the accuracies of the SEEs and SEEDs produced results that were similar to those of previous studies of equating error estimation (Jarjoura & Kolen, 1985; Liou et al., 1997; Liou & Cheng, 1995; Lord, 1982). Unlike equating function accuracy, SEE and SEED accuracy is not very dependent on the accuracy of the presmoothing model: Stronger smoothing models that were incorrect (M221) resulted in low and extremely accurate variability estimates, while the accuracies of the raw SEEs were subject to small-sample and/or sparse data problems. Theoretical standard errors based on the delta method are fairly accurate except in conditions where overall samples are small and the data at parts of the score ranges are sparse. The tendency of raw post-stratification SEEs to underestimate actual variability has been found in previous studies (Jarjoura & Kolen, 1985; Liou & Cheng, 1995). The evaluation of the accuracy of the SEEDs has not been previously considered, and the results show that SEED accuracies are generally high and similar to SEE accuracies.

The implications of presmoothing correctness for equating function and SEE accuracy were very similar for the chained and post-stratification methods. The extremely simple M221 model’s relatively large bias and small variability were visible in both equating methods. Similarly, the over-parameterized M661 model’s small bias and large variability was also visible in both equating methods. The chained kernel method with the M661 presmoothing model had more marked tail-bias than other methods and also had more unrealistically large SEE estimates. These problems were likely due to this method’s use of the extreme and sparse scores of the four (over-parameterized) marginal distributions, combined with the kernel method’s tendency to make overly fine distinctions in the extreme and sparse parts of the score distribution.

Comparisons of Kernel and Traditional Equipercentile Continuizations

One of the focuses of this study was investigating a likely interaction between a given presmoothing model and the traditional percentile-rank and kernel continuization methods. (Formal discussions of each continuization method are included in Holland & Thayer, 1989, and von Davier et al., 2004.) This section compares how these methods addressed data that had been presmoothed in particular ways, as well as their final equating variability estimates.

Continuization score distributions. Figures 35–38 illustrate how the M221, M661, MP, and raw distributions of the A_Q distribution generated in the first replication of the N = 200 samples were continuized with the kernel and percentile-rank methods. When data have been strongly presmoothed, such as with the M221 or M661 models (Figures 35 and 36), the differences in the continuization methods are barely visible in the plots, except at the most abrupt changes in the distributions. The selected kernel bandwidths (.539 and 1.018) are relatively small, resulting in kernel continuizations that mostly utilize the nearest scores’ data, as does the percentile-rank continuization method. As the teeth are incorporated into the smoothing models (MP, Figure 37) and as the raw data are considered (Figure 38), the differences in the percentile-rank and kernel continuization methods become much more visible. The discrete and piece-wise nature of the fixed-interval percentile-rank method is clearly visible. The kernel continuization methods are based on larger bandwidths (1.542 and 2.694) that utilize much wider score intervals than the data that were presmoothed with a stronger model. Unlike the percentile-rank continuizations, the resulting kernel continuizations are always extremely smooth, staying very close to the majority of nearby score probabilities as the discrete distributions fluctuated.

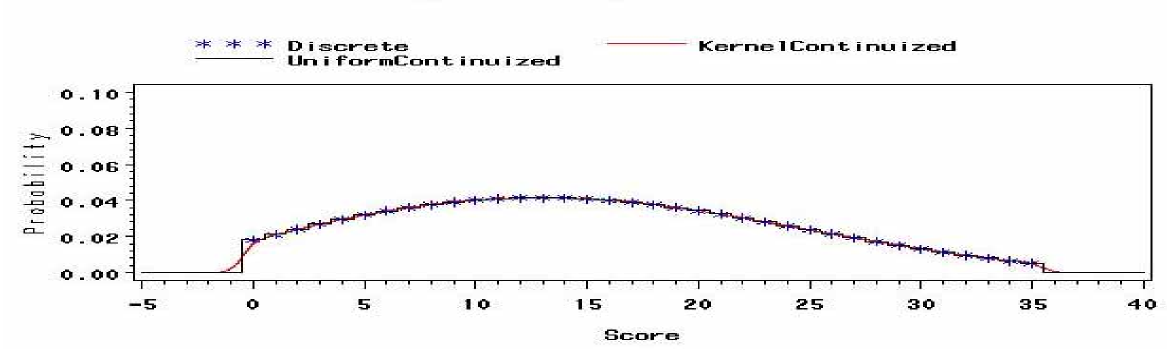

Figure 35. **Discrete and continuized probabilities, M221.**

Note. Data = AQ, H = 0.5392103959.

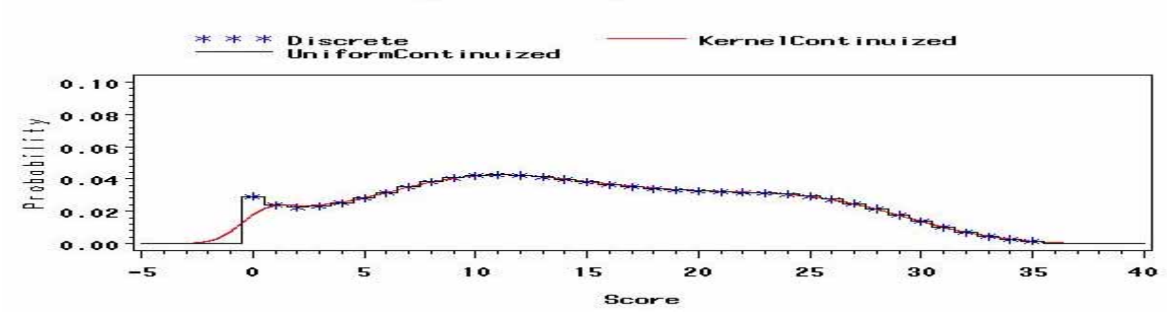

Figure 36. **Discrete and continuized probabilities, M661.**

Note. Data = AQ, H = 1.0176032685.

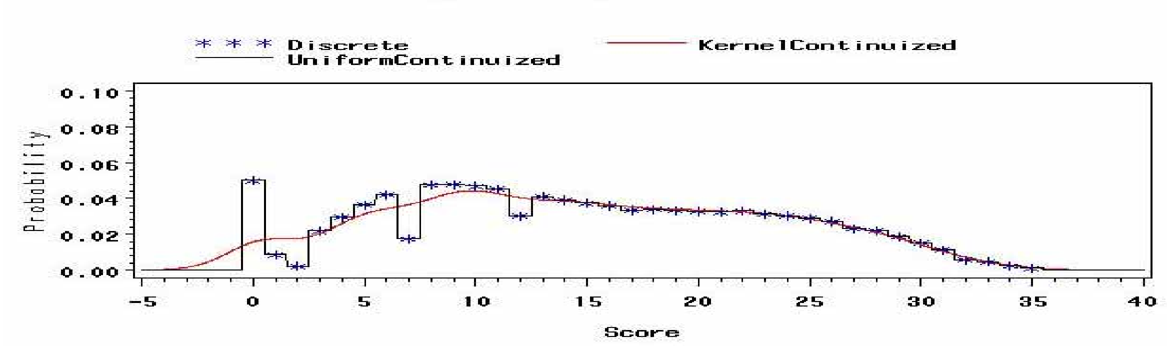

Figure 37. **Discrete and continuized probabilities, MP.**

Note. Data = AQ, H = 1.5418304658.

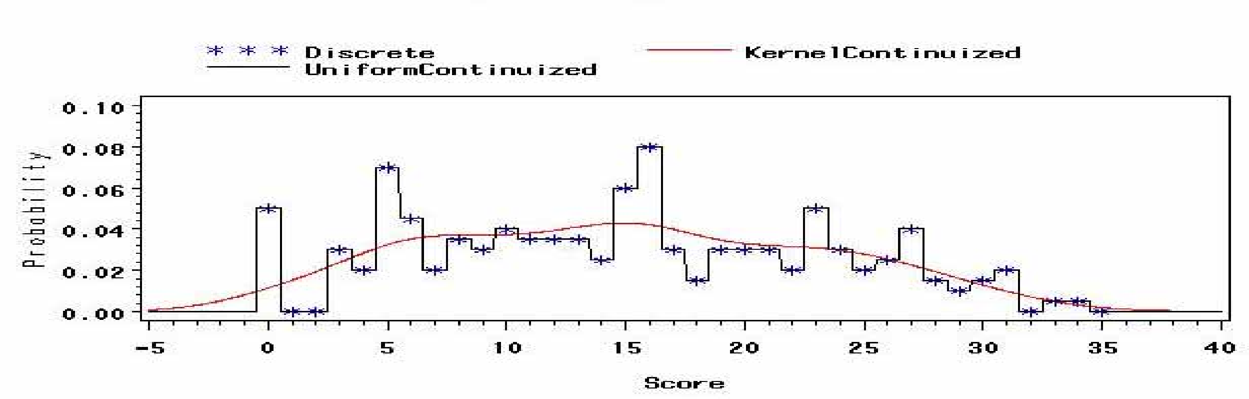

Figure 38. **Discrete and continuized probabilities, raw.**

Note. Data = AQ, H = 2.6936097909.

Raw continuization and variability. Using the kernel and percentile-rank continuizations with given degrees of presmoothing have different implications for the accuracies of the equating functions and the SEE estimates. The kernel continuization method utilized a continuization bandwidth that was based on the extent of roughness in the presmoothed data. Simpler presmoothing models resulted in the selection of smaller continuization bandwidths, while complicated presmoothing and raw data resulted in the selection of larger continuization bandwidths (Tables 4-7). The roughness in the data due to the teeth, the lumps at zero, and raw fluctuation was smoothed out to a greater extent when the kernel continuization method was used than when the percentile-rank continuization method was used. The result was that kernel SEEs were smaller than traditional equipercentile SEEs, supporting previous work (Liou et al., 1997).

The differences in the standard errors are shown in one individual replication of the chained (Figure 39) and post-stratification (Figure 40) equating functions for sample size combination $N_P = N_Q = 200$. The series of raw equipercentile standard errors for single replications fluctuate greatly while the raw kernel standard errors are stable. It is clear that when applied to raw data, the kernel continuization functions as a kind of post-smoother for SEE series that makes the series of raw equipercentile standard errors more reasonable (Harris & Kolen, 1986).

Raw continuization and SEEDs. An interesting aspect of this study's results that can be used to further compare the kernel and uniform continuization approaches is the failure of the average raw equipercentile chained post-stratification SEED to accurately approximate the actual variability of equated score differences (Figure 30). The difference in degree of raw continuization for the kernel and traditional equipercentile functions is one potential explanation for the raw equipercentile SEEDs in Figure 30.

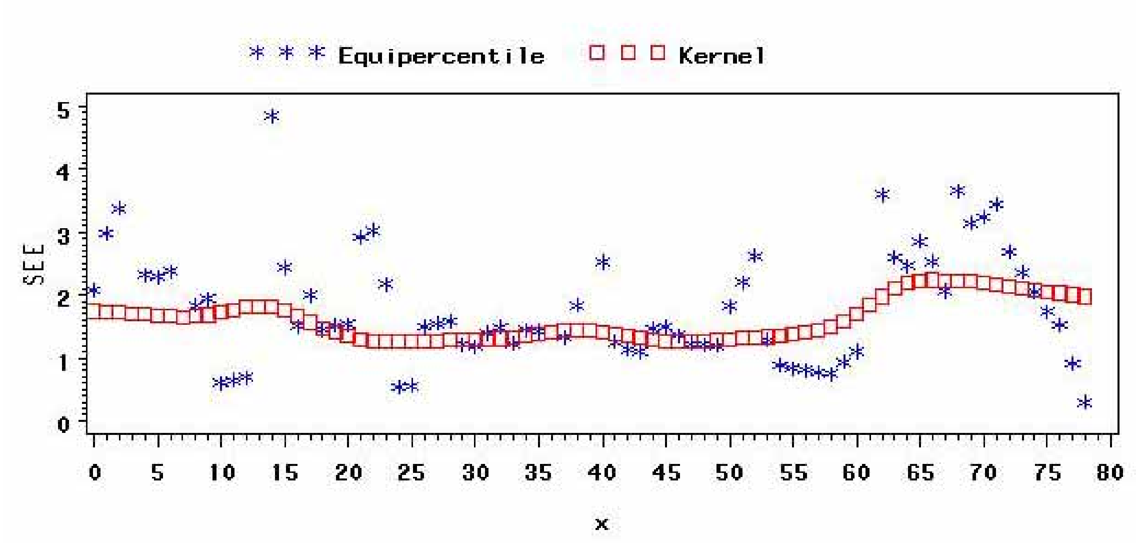

Figure 39. **Raw chained SEEs.**

Note. *Np* = *Nq* = 200.

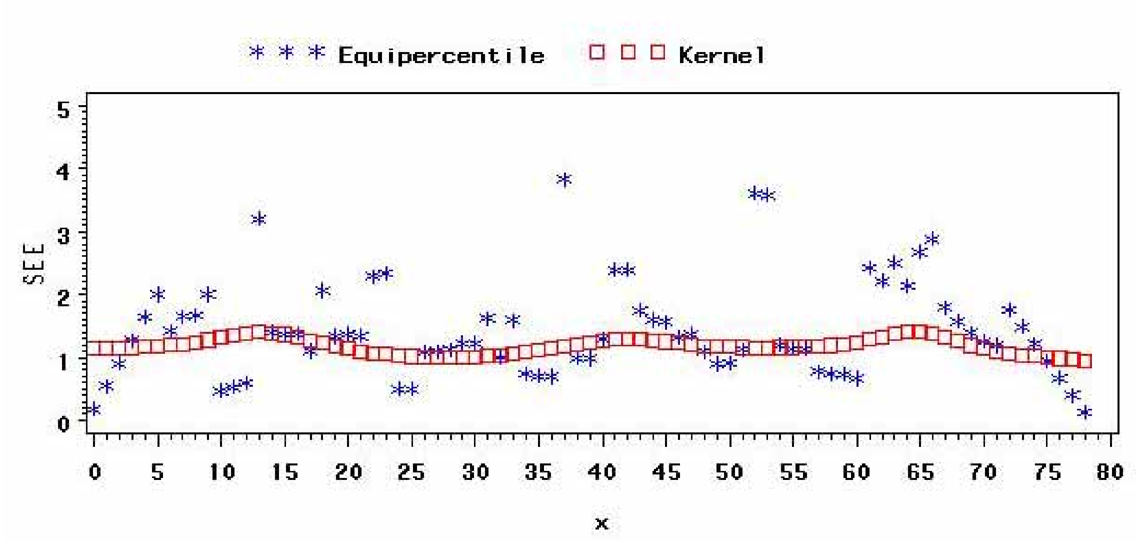

Figure 40. **Raw post-stratification SEEs.**

Note. *Np* = *Nq* = 200.

Two follow-up analyses were conducted on raw kernel and equipercentile SEEDs. One question evaluated was whether the kernel SEED could be made to overestimate actual variability when its degree of continuization was made to be more similar to that of the traditional equipercentile method. Figure 41 plots the kernel SEEDs and actual standard deviations of differences where the raw kernel functions were based on very small continuization bandwidths (.3 for N = 1,000 and .4 for N = 200). In comparison to the raw kernel SEEDs based on larger bandwidths (Figure 34), Figure 41 shows that some overestimation resulted from the smaller bandwidths, particularly for the tails of the $N_P = N_Q = 200$ SEEDs. The overestimation, although not of the same magnitude as the overestimation of raw equipercentile SEEDs (Figure 30), is clearly a function of the extent to which the raw data are strongly or weakly continuized.

A second question was whether the raw equipercentile SEEDs' overestimation problems could be controlled by utilizing wider ranges of the score distribution data. To evaluate this question, the procedure used in this study for dealing with scores with zero frequencies in raw equipercentile equating was to average the score frequencies that were less than one (Moses & Holland, 2007). For this follow-up, the data from scores with less than 1.5% (15 for N = 1,000 and 3 for N = 200) were averaged together. This stronger degree of frequency averaging resulted in the raw equipercentile SEED plot displayed in Figure 42, which shows more accurate SEEDs than those in Figure 30.

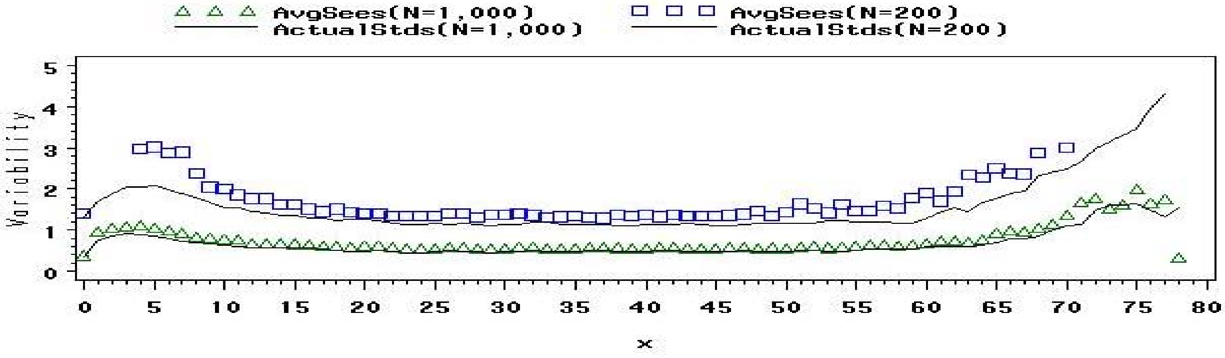

***Figure 41.* Chained and post-stratification kernel SEED, raw.**

Note. Low bandwidths (.3 for N = 1,000; .4 for N = 200).

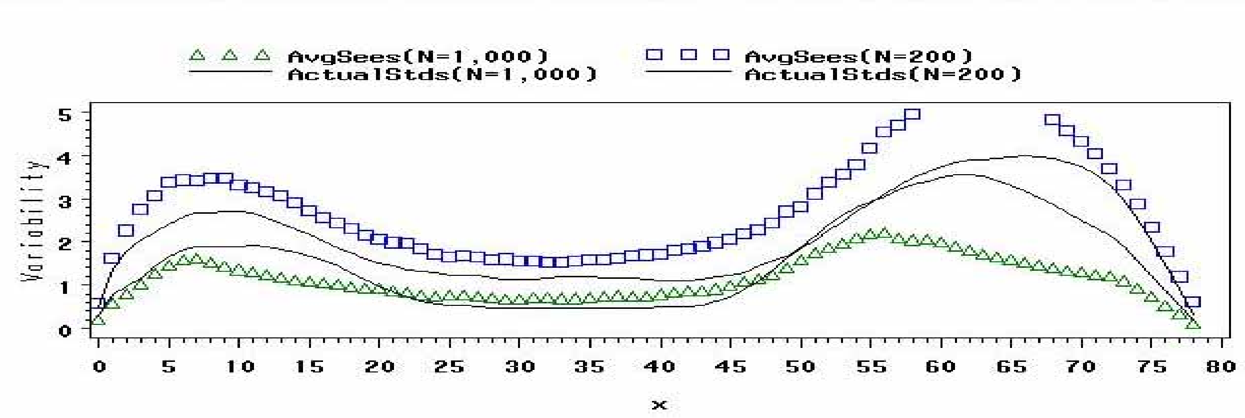

Figure 42. **Chained and post-stratification equipercentile SEED, raw.**

Note. High frequency averaging (minimum frequency = 15 for *N* = 1,000 and 3 for *N* = 200).

Conclusion

Two of the most important conclusions from the results of this study's varying the degree of loglinear presmoothing on NEAT equating results concern equating bias and the interaction of presmoothing and continuization. In regard to the former, the degree of presmoothing has a direct relationship to equating bias, but one such that only the simplest models produce bias levels that would have practical effects on equating functions. Reasonably complex models (ones preserving at least three moments in the marginal distributions, which is probably warranted in the majority of presmoothing situations) probably produce negligibly biased equating functions across many equating situations. In addition, extensive model-search evaluations (e.g., Holland & Thayer, 2000; von Davier et al., 2004) to obtain the closest possible reflection of an infinitely complex population distribution may not be needed to produce a negligibly biased equating function that is less variable than when estimated using raw data.

The results from comparing how degrees of presmoothing work with continuization methods show that the kernel continuization method has some potentially important advantages over the percentile-rank method. In terms of raw, unsmoothed data, the kernel method can continuize even when there are scores with zero frequencies, while the percentile-rank method requires additional procedures to handle such scores (Kolen & Brennan, 2004). When the roughness of rounded formula-scored distributions is preserved in the smoothing model, the kernel method's data-adaptive use of larger continuization bandwidths avoids some small perturbations in the bias and SEE series that are inherent when the percentile-rank method is used. The tradeoff for the kernel method’s flexibility is that the method can introduce some bias, such as was found for the chained equating method. Across all of the degrees of presmoothing and sample sizes considered in this study, the kernel method’s bias appears to be small in relation to its flexibility, making the method especially well suited for considering how equating results are affected across degrees of presmoothing.

References

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15(4),* 391–408.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Technical Report 94-4). Iowa City, IA: ACT.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10(1),* 35–43.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Rep. No. ETS RR-87-31). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Research Rep. No. RR-89-07). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–83.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10,* 143–160.

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

Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20,* 259–286.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21(4),* 349–369.

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

Lord, F. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics, 7*(3), 165–174.

Moses, T., & Holland, P. (2007). *Notes on a general framework for observed score equating.* Manuscript in preparation.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). New York: Cambridge University Press.

Skaggs, G. (2004, April). *Passing score stability when equating with very small samples.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Sinharay , S., & Holland, P. W. (2006). *Choice of anchor test in equating* (ETS Research Rep. No. RR-06-35). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer-Verlag.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2006, April). *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 the National Council on Measurement in Education, San Francisco CA.