Research Report

The Effects of Different Types of Anchor Tests on Observed Score Equating

Jinghua Liu

Sandip Sinharay

Paul W. Holland

Miriam Feigenbaum

Edward Curley

December 2009

ETS RR-09-41

The Effects of Different Types of Anchor Tests on Observed Score Equating

Jinghua Liu, Sandip Sinharay, Paul W. Holland, Miriam Feigenbaum, and Edward Curley
ETS, Princeton, New Jersey

December 2009

Abstract

This study explores the use of a different type of anchor, a *midi anchor*, that has a smaller spread of item difficulties than the tests to be equated, and then contrasts its use with the use of a *mini anchor*. The impact of different anchors on observed score equating were evaluated and compared with respect to systematic error (bias), random equating error (standard errors of equating), and total equating error (RMSE) using empirical data. The results show that the midi anchor generally produces more accurate equating results than the mini anchor—with a few exceptions. Our findings suggest that a midi anchor test would be preferred to a mini anchor test if equating accuracy at the top and at the bottom of the score scale is not a primary concern.

Key words: Anchor test equating, middle difficult anchor, mini anchor

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Introduction

The nonequivalent-groups anchor test (NEAT) design is often employed for test score equating. In the NEAT design, population *P* takes form *X*, population *Q* takes form *Y*, and both populations take the same anchor test *A*. The anchor test is used to control for differential ability by quantifying the ability difference between the samples of test- anchor should be a mini version of the tests to be equated. Kolen and Brennan (2004) suggested that anchor should be built to the same specifications as the total tests, in both content and takers from *P* and *Q*. It is critical that the anchor reflects the differences between the groups taking the new form and the old form (Livingston, 2004) and it is commonly advised that an statistical characteristics, in order to reflect the group differences accurately (p. 9; p. 271). Livingston recommended choosing common items that resemble the full test in content and format, and that represent the full range of difficulty (p. 38). Several other experts recommended anchor tests that are mini versions of the total tests (Angoff, 1984; Dorans, Kubiak, & Melican, 1998; Holland, Dorans, & Peterson, 2006; Peterson, Kolen & Hoover, 1989; von Davier, Holland, & Thayer, 2004.).

Sinharay and Holland (2006; 2007), however, challenged this traditional view of the need for a mini anchor test. While they acknowledged the importance of an anchor being content representative and having the same mean difficulty as the tests to be equated, they questioned the restriction that an anchor test needs to have the same spread of difficulty as the total test. They considered anchor tests that are content representative and have the same mean item difficulty as the tests to be equated, but have a spread of item difficulties less than of the total test. These anchors are referred as *midi anchors*. Using simulated data and pseudo data, Sinharay and Holland showed that midi anchors perform as well as mini anchors with respect to equating bias, equating standard error, and root mean squared error in equating. They recommended comparing the equating performances of midi anchors and mini anchors using more operational data (Sinharay & Holland, 2007, p. 273).

The current study can be seen as a follow-up to Sinharay and Holland (2007). In this study, we use real data and investigate two different types of anchors—a mini anchor and a midi anchor—and their impact on observed score equating. One feature that distinguishes this research from the previous studies is the use of operational data rather than simulated data. In addition, the criterion used in this study is also different. In Sinharay and Holland (2007), two different populations were combined to form a single group, and a single-group equipercentile equating was used as the true equating function. Because the two populations differed in ability (about a quarter of the raw score standard deviation in size), combining them to create a single population could introduce errors into the equating function. In contrast, the current study computes the criterion equating function based on two truly equivalent groups with ability difference near zero in raw score standard deviation units.

Methodology

Equating Design

In this study, we used data from the verbal section of old SAT® I forms. In a typical SAT I administration, a new test was administered in different forms. One was called the *original order form*, and the other was called the *scrambled form*. The two forms contained the same operational items, but the order of the separately timed sections in the forms was different. The two forms were spiraled during an administration. After the equatings were conducted on the original order form and a conversion was produced, the same conversion was usually applied to the scrambled form, unless a section order effect was detected. In this study, we treated the scrambled form as the new form X and the original order form as the old form Y. Both forms contained 78 operational items, and form X was equated to form Y through an external anchor A (the choice of A will be discussed shortly). We used this study design to get a group with similar abilities so that equivalent groups design equating could be used as the equating criterion and to study the performance of the anchor tests in a situation where there were no test form differences.

Three different anchors were used in this study. The first one was the operational anchor administered with both forms. This anchor had 35 items and was a mini version of the total test. We called this anchor an *intact anchor* to distinguish it from another mini anchor. Based on the intact anchor, two other shorter anchors were built. One was referred to as a *mini anchor*, which had the same mean and standard deviation of item difficulties as the total test. The other was referred to as a *midi anchor*. Its mean item difficulty was the same as that of the total test, but its standard deviation of item difficulty was only half of that of the total test. Note that strictly speaking, this is actually a *semi–midi anchor* according to the terminology of Sinharay and Holland (2007). Each of the shorter anchors was content representative and contained 20 items, with 9 of the items overlapping across the two shorter anchors. Therefore, form X could be equated to form *Y* through three different anchors: the intact anchor, the mini anchor, and the midi anchor.

Equating Data Sets

To examine whether the results were replicable, we ran equatings and analyses on two different data sets. Each data set came from a different administration, and each data set consisted of data from an original order form, a scrambled form, and an intact anchor. We used symbols *X1* and *Y1* to designate the first data set and symbols *X2* and *Y2* to designate the second data set.

Table 1 presents the mean and standard deviation of the item difficulty indexes for the total test and the three anchors for both data sets. The item difficulty index used is the delta index, which is based on the delta scale. The delta scale is a standard scale converted from *p*-values, where (1 – *p*) is first converted to a normalized *z*-score and then linearly transformed to a scale with a mean of 13 and a standard deviation of 4: $Delta = 13 + 4 \times z$. Delta values can range from 4 to 22, corresponding to *p*-values of 99% and 1%, respectively. Deltas are inversely related to *p*-values.

Table 1

Mean and Standard Deviation of the Item Difficulty of Total Tests and Three Anchors

	Total test	Intact anchor	Mini anchor	Midi anchor
Data set 1				
N of items	78	35	20	20
Mean difficulty	11.40	11.37	11.32	11.34
SD of difficulty	2.23	2.36	2.28	1.29
Data set 2				
N of items	78	35	20	20
Mean difficulty	11.40	11.39	11.44	11.42
SD of difficulty	2.23	2.33	2.23	1.10

For both data sets, the mean delta of each anchor was very close to that of the total test, but the standard deviation of the delta in the midi anchor was only half that of the total test while the intact anchor and the mini anchor had similar standard deviations of the deltas as the total test.

Equating Samples

Forms *X1* and *Y1* were spiraled in one old SAT I administration. Given the large number of test-takers (more than 200,000 test-takers took forms *X1* and *Y1,* respectively) and the spiraling procedure used in the administration, the group *P1,* which took form *X1,* and the group *Q1,* which took form *Y1,* were deemed to be very similar. Forms *X2* and *Y2* were spiraled in a different administration. Approximately 100,000 test-takers took forms *X2* and *Y2*, respectively. The group *P2*, which took form *X2*, was deemed to be very similar to the group *Q2*, which took form *Y2*.

It is well known that differences in the abilities of the old and new form samples affect the quality of equating (e.g., Kolen, 1990; Sinharay & Holland, 2007). To compare the anchor tests under different levels of equating sample ability differences, we identified five subgroups (SG) with different abilities based on the examinee's demographic information. The subgroups were referred to as SG1, SG2, SG3, SG4, and SG5. We then constructed four pairs of equating samples with varied ability differences: total group to total group (very similar in ability), SG1 to SG2 (moderately similar in ability), SG3 to SG5 (moderately dissimilar in ability), and SG4 to SG5 (very dissimilar in ability). In each pair, we used one member as the new form sample and the other as the old form sample. For example, the total group taking form *X1* and the total group taking form *Y1* were used to equate *X1* to *Y1* to emulate an equating with very similar old form and new form samples; test-takers belonging to SG1 and taking form *X1,*and those belonging to SG2 and taking form *Y1* are used to equate *X1* to *Y1* to emulate an equating with moderately similar samples. For each pair, we performed equating using anchor tests.

The raw score descriptive statistics for equating samples in data set 1 are presented in Tables 2 to 5, for total group to total group, SG1 to SG2, SG3 to SG5, and SG4 to SG5, respectively. Tables 2 to 5 also list the standardized mean difference between the new form sample of the test-takers and the old form sample of the test-takers on each of the three anchors. A negative sign indicates that the new group was less able than the old group. Data in Table 2 show that the two total groups were very similar to each other, with the standardized mean difference on the anchor test close to zero across the three different anchors. So this equating used a pair of samples that were very similar. The similarity of the two total groups is also shown in the percentages of means (around 50% for each group) and standard deviations (around 25%)

Table 2

Descriptive Statistics of Raw Scores in the Total-Group-to-Total-Group Equating: X1 to Y1

	New form sample: total group				Old form sample: total group			
	Test *X1*	Anchor *A1*			Anchor *A1*			Test *Y1*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	11,302	11,302	11,302	11,302	11,361	11,361	11,361	11,361
# of items	78	35	20	20	35	20	20	78
Mean	39.13	17.90	10.26	10.35	17.97	10.28	10.36	39.78
Mean % correct	50%	51%	51%	52%	51%	51%	52%	51%
SD	17.10	8.16	4.85	5.49	8.29	4.90	5.52	17.21
SD % correct	22%	23%	24%	27%	24%	25%	28%	22%
Skewness	-0.10	-0.10	-0.12	-0.11	-0.10	-0.14	-0.10	-0.13
Kurtosis	2.27	2.29	2.34	2.14	2.24	2.33	2.08	2.26
Reliability	0.92	0.85	0.77	0.79	0.86	0.78	0.80	0.93
Correlation		0.88	0.85	0.85	0.89	0.85	0.86	

	Mini	Midi
Std. mean difference on the anchor (new–old)	-0.004	-0.002
Ratio of variance on the anchor (new/old)	0.980	0.989

Table 3

Descriptive Statistics of Raw Scores in the SG1-to-SG2 Equating: X1 to Y1

	New form sample: SG1				Old form sample: SG2			
	Test *X1*	Anchor *A1*			Anchor *A1*			Test *Y1*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	6,549	6,549	6,549	6,549	4,901	4,901	4,901	4,901
# of items	78	35	20	20	35	20	20	78
Mean	39.14	18.00	10.23	10.46	17.98	10.44	10.30	40.20
Mean % correct	50%	51%	51%	52%	51%	52%	52%	52%
SD	16.97	8.11	4.81	5.46	8.34	4.93	5.57	17.26
SD % correct	22%	23%	24%	27%	24%	25%	28%	22%
Skewness	-0.11	-0.10	-0.11	-0.13	-0.11	-0.18	-0.10	-0.16
Kurtosis	2.27	2.30	2.35	2.16	2.24	2.35	2.06	2.25
Reliability	0.92	0.84	0.76	0.78	0.86	0.78	0.80	0.93
Correlation		0.88	0.84	0.85	0.89	0.85	0.86	

	Mini	Midi
Std. mean difference on the anchor (new–old)	-0.043	0.029
Ratio of variance on the anchor (new/old)	0.952	0.961

Table 4

Descriptive Statistics of Raw Scores in the SG3-to-SG5 Equating: X1 to Y1

	New form sample: SG3				Old form sample: SG5			
	Test *X1*	Anchor *A1*			Anchor *A1*			Test *Y1*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	973	973	973	973	6,291	6,291	6,291	6,291
# of items	78	35	20	20	35	20	20	78
Mean	36.55	16.56	9.66	9.19	19.02	10.86	11.10	42.18
Mean % correct	47%	47%	48%	46%	54%	54%	56%	54%
SD	18.51	8.95	5.31	5.88	7.85	4.68	5.25	16.00
SD % correct	24%	26%	27%	29%	22%	23%	26%	21%
Skewness	-0.03	0.01	-0.02	0.08	-0.14	-0.19	-0.16	-0.15
Kurtosis	2.07	2.05	2.07	2.01	2.31	2.42	2.14	2.30
Reliability	0.94	0.87	0.80	0.81	0.85	0.76	0.78	0.92
Correlation		0.90	0.87	0.87	0.88	0.84	0.85	
					Mini	Midi		
Std. mean difference on the anchor (new–old)					-0.252	-0.358		
Ratio of variance on the anchor (new/old)					1.287	1.254		

Table 5

Descriptive Statistics of Raw Scores in the SG4-to-SG5 Equating: X1 to Y1

	New form sample: SG4				Old form sample: SG5			
	Test *X1*	Anchor *A1*			Anchor *A1*			Test *Y1*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	788	788	788	788	6,291	6,291	6,291	6,291
# of items	78	35	20	20	35	20	20	78
Mean	28.14	13.26	7.53	7.18	19.02	10.86	11.10	42.18
Mean % correct	36%	38%	38%	36%	54%	54%	56%	54%
SD	16.50	7.81	4.68	5.30	7.85	4.68	5.25	16.00
SD % correct	21%	22%	23%	27%	22%	23%	26%	21%
Skewness	0.34	0.39	0.37	0.40	-0.14	-0.19	-0.16	-0.10
Kurtosis	2.28	2.50	2.56	2.44	2.31	2.42	2.14	2.27
Reliability	0.92	0.83	0.77	0.75	0.85	0.76	0.78	0.92
Correlation		0.87	0.84	0.84	0.88	0.84	0.85	
					Mini	Midi		
Std. mean difference on the anchor (new–old)					-0.712	-0.746		
Ratio of variance on the anchor (new/old)					1.000	1.019		

of the total number of items. The SG1-to-SG2 differences were a little larger when compared to the total-to-total group differences but still less than .05 of a standard deviation unit (Table 3). In contrast, the SG3-to-SG5 comparison (see Table 4) indicated that the two subgroups differed moderately, with the standardized mean differences close to 0.3. The percentages of the means out of the total number of items were near 46% to 48% for SG3 but around 54% to 56% for SG5. Finally, the SG4-to-SG5 comparison (Table 5) exhibited the largest ability differences, around 0.7. The mean percentages were only below 40% for SG4.

Tables 2 to 5 also present the reliability and the correlation between the scores on the anchor and those on the total test. As shown in Table 2, the 35-item intact anchor had higher reliability and higher anchor-to-total test correlation than those of the 20-item mini anchor and the 20-item midi anchor. When we compared the mini anchor with the midi anchor, the midi anchor correlated with the test to be equated either slightly higher than or as same as the mini anchor did. The same pattern was observed in the other equatings: SG1 to SG2, SG3 to SG5, and SG4 to SG5 (see Tables 3 to 5). This finding is consistent with Sinharay and Holland (2006; 2007).

Tables 6 to 9 contain raw score descriptive statistics for equating samples in data set, for total group to total group, SG1 to SG2, SG3 to SG5, and SG4 to SG5, respectively. The groups exhibited a similar pattern as the groups in data set 1: the two total groups were very similar to each other, and SG1 and SG2 were moderately similar. The subgroups SG3 and SG5 differed moderately in abilities, with the standardized mean differences around 0.3, whereas SG4 and SG5 showed the largest ability difference, around 0.8.

Table 6 shows that the midi anchor correlated with the total test slightly higher than the mini anchor did, both in the new form samples (0.84 vs. 0.83) and in the old form samples (0.85 vs. 0.84). The same pattern can be observed in the other equatings: SG1 to SG2, SG3 to SG5, and SG4 to SG5 (Tables 7 to 9).

Equating Methods

Two types of equating methods were used: chained equating (CE) and poststratification equating (PSE). Chained equating uses the anchor as part of a chain: first link X to A on P, and then to link A to Y on Q. The two linking functions are then composed to map X to Y through A. Two CE methods, the chained linear method and the chained equipercentile method, were both conducted in this study.

Table 6

Descriptive Statistics of Raw Scores in the Total-Group-to-Total-Group Equating: X2 to Y2

	New form sample: total group				Old form sample: total group			
	Test *X2*	Anchor *A2*			Anchor *A2*			Test *Y2*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	7,359	7,359	7,359	7,359	7,528	7,528	7,528	7,528
# of items	78	35	20	20	35	20	20	78
Mean	34.51	16.17	9.32	9.53	16.06	9.17	9.49	34.96
Mean % correct	44%	46%	47%	48%	46%	46%	47%	45%
SD	17.59	7.68	4.61	5.16	7.82	4.66	5.28	17.82
SD % correct	23%	22%	23%	26%	22%	23%	26%	23%
Skewness	0.17	0.04	0.00	-0.07	0.01	0.02	-0.11	0.13
Kurtosis	2.25	2.49	2.47	2.28	2.41	2.43	2.25	2.22
Reliability	0.92	0.84	0.75	0.77	0.84	0.76	0.78	0.93
Correlation		0.88	0.83	0.84	0.88	0.84	0.85	

	Mini	Midi
Std. mean difference on the anchor (new–old)	0.032	0.008
Ratio of variance on the anchor (new/old)	0.979	0.955

Table 7

Descriptive Statistics of Raw Scores in the SG1-to-SG2 Equating: X2 to Y2

	New form sample: SG1				Old form sample: SG2			
	Test *X2*	Anchor *A2*			Anchor *A2*			Test *Y2*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	4,151	4,151	4,151	4,151	3,383	3,383	3,383	3,383
# of items	78	35	20	20	35	20	20	78
Mean	33.87	15.86	9.08	9.37	16.32	9.39	9.63	35.57
Mean % correct	43%	45%	45%	47%	47%	47%	48%	46%
SD	17.54	7.66	4.56	5.16	7.86	4.72	5.27	17.87
SD % correct	22%	22%	23%	26%	22%	24%	26%	23%
Skewness	0.19	0.06	0.02	-0.05	-0.04	-0.02	-0.16	0.08
Kurtosis	2.25	2.49	2.48	2.29	2.37	2.41	2.24	2.17
Reliability	0.93	0.84	0.75	0.77	0.84	0.77	0.78	0.93
Correlation		0.88	0.83	0.84	0.88	0.85	0.85	

	Mini	Midi
Std. mean difference on the anchor (new–old)	-0.067	-0.050
Ratio of variance on the anchor (new/old)	0.933	0.959

Table 8

Descriptive Statistics of Raw Scores in the SG3-to-SG5 Equating: X2 to Y2

	New form sample: SG3				Old form sample: SG5			
	Test *X2*	Anchor *A2*			Anchor *A2*			Test *Y2*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	559	559	559	559	3,802	3,802	3,802	3,802
# of items	78	35	20	20	35	20	20	78
Mean	33.35	15.66	9.01	9.19	17.66	10.09	10.58	38.50
Mean % correct	43%	45%	45%	46%	50%	50%	53%	49%
SD	18.70	8.33	4.95	5.48	7.16	4.31	4.82	16.44
SD % correct	24%	24%	25%	27%	20%	22%	24%	21%
Skewness	0.21	0.15	0.00	0.07	-0.05	-0.05	-0.21	0.05
Kurtosis	2.11	2.23	2.17	2.09	2.59	2.59	2.43	2.29
Reliability	0.93	0.84	0.77	0.78	0.82	0.73	0.75	0.92
Correlation		0.88	0.85	0.85	0.87	0.82	0.83	

	Mini	Midi
Std. mean difference on the anchor (new–old)	-0.246	-0.283
Ratio of variance on the anchor (new/old)	1.319	1.293

Table 9

Descriptive Statistics of Raw Scores in the SG4-to-SG5 Equating: X2 to Y2

	New form sample: SG4				Old form sample: SG5			
	Test *X2*	Anchor *A2*			Anchor *A2*			Test *Y2*
		Intact	Mini	Midi	Intact	Mini	Midi	
N	786	786	786	786	3,802	3,802	3,802	3,802
# of items	78	35	20	20	35	20	20	78
Mean	23.04	11.82	6.79	6.69	17.66	10.09	10.58	38.50
Mean % correct	30%	34%	34%	33%	50%	50%	53%	49%
SD	14.97	6.72	4.09	4.65	7.16	4.31	4.82	16.44
SD % correct	19%	19%	20%	23%	20%	22%	24%	21%
Skewness	0.72	0.35	0.26	0.32	-0.05	-0.05	-0.21	0.05
Kurtosis	3.21	2.95	2.91	2.61	2.59	2.59	2.43	2.29
Reliability	0.90	0.79	0.68	0.71	0.82	0.73	0.75	0.92
Correlation		0.84	0.78	0.80	0.87	0.82	0.83	

	Mini	Midi
Std. mean difference on the anchor (new–old)	-0.772	-0.812
Ratio of variance on the anchor (new/old)	0.901	0.931

Poststratification equating uses the anchor test *A* to estimate the distribution of *X* on *Q* and the distribution of *Y* on *P*. It assumes that the conditional distribution of *X* given *A* and the conditional distribution of *Y* given *A* are population invariant and then poststratifies the distributions of both *X* and *Y* on a target population *T* (synthetic population of *P* and *Q*). The PSE methods used in this study were the Tucker (linear) method and the frequency estimation (nonlinear) method.

Equating Criterion

As mentioned above, the large number of test-takers and the spiraling procedure usually yield equivalent groups in each SAT administration. Therefore, we conducted linear equating and equipercentile equating to equate the scrambled form *X1* to the original order form *Y1* (and *X2* to *Y2)* through an equivalent groups (EG) design. Since the forms to be equated are essentially the same form but just in different section orders, the linear conversion based on 205,793 test-takers on the new form *X1* and 210,681 test-takers on the old form *Y1* was very close to the identity conversion, with a = 1.0001 and b = .0581 based on the raw score scale. As expected, when the linear equating conversion was compared to its curvilinear analogue, the differences were negligible. Hence, the linear conversion was used as the criterion equating function. Similarly, the linear conversion with a = 0.9989 and b = 0.2696 based on 99,653 test-takers taking new form *X2* and 102,568 test-takers taking form *Y2* was used as the criterion equating function for the second data set.

Discrepancy Indices

Bias. Bias measures the systematic error in equating. The bias at each score point can be calculated by

$$Bias\left[\hat{e}_Y(x_i)\right] = \hat{e}_Y(x_i) - e_Y(x_i)\,, \qquad (1)$$

where $e_Y(x_i)$ denotes the criterion of the population equating function at score point x_i, and $\hat{e}_Y(x_i)$ is the sample equating function derived from the anchor test equating. To summarize the differences over all score points, we calculate the weighted absolute bias given by

$$Weighted\ \ Absolute\ Bias\left[\hat{e}_Y(x)\right] = \frac{1}{N}\sum f_{x_i}\left|Bias\left[\hat{e}_Y(x_i)\right]\right|\ . \qquad (2)$$

N is the total number of test-takers in the new form equating sample, and f_{x_i} is the frequency at score point x_i in the new form group.

To evaluate the magnitude of the bias, we used the notion of *score difference that matters* (DTM) as an effect size criterion (Dorans & Feigenbaum, 1994; Holland & Dorans, 2006). We defined 0.5 (in absolute value), which was half of the raw score scale unit, as one DTM, because we were equating on the raw score scale. A difference of 0.5 or more usually led to a different equated score.

Standard errors of equating (SEE). The SEE measures random error in equating due to sampling variability. The conditional SEE at score point x_i is defined as the square root of the error variance (Kolen & Brennan, 2004),

$$SEE\left[\hat{e}_Y(x_i)\right] = \sqrt{\mathrm{var}[\hat{e}_Y(x_i)]}\,. \tag{3}$$

The calculation of the conditional SEE in this study was based on the delta method (Kolen & Brennan). Similarly, the weighted average SEE is defined as

$$Weighted\ SEE\left[\hat{e}_Y(x)\right] = \frac{1}{N}\sum f_{x_i}\left\{SEE\left[\hat{e}_Y(x_i)\right]\right\}. \tag{4}$$

Root mean squared error (RMSE). Considering both random error and systematic error, a root mean squared error (RMSE) can be calculated at each score point as

$$RMSE\left[\hat{e}_Y(x_i)\right] = \sqrt{\left\{Bias\left[\hat{e}_Y(x_i)\right]\right\}^2 + \left\{SEE\left[\hat{e}_Y(x_i)\right]\right\}^2}\,. \tag{5}$$

To get a measure of overall error across all the score points, the weighted RMSE is defined as

$$Weighted\ RMSE\left[\hat{e}_Y(x)\right] = \frac{1}{N}\sum f_{x_i}\left\{RMSE\left[\hat{e}_Y(x_i)\right]\right\}. \tag{6}$$

Results

Bias

Figures 1 to 8 present the bias plots of each equating function compared to the criterion equating. The horizontal axis represents raw score points on form *X1 or X2*, and the vertical axis indicates the magnitude of the bias.

Total-group-to-total-group equating. Figure 1 shows the results of the total-group-to-total-group equating in data set 1. As can be seen from Figure 1, in general, the three different types of anchor tests performed very similarly, with the three bias lines/curves close to each other. The midi anchor test did slightly better at the lower range of the scale, whereas the intact anchor did slightly better at the upper range of the scale. Nevertheless, the deviations from the criterion were smaller than or very close to the DTM across most of the scale range (with a few exceptions).

The bias plot of total-group-to-total-group equating for data set 2 is shown in Figure 2. The midi anchor outperformed the mini anchor across the entire score range for Tucker, frequency estimation, and chained linear equatings. In the chained equipercentile equating, the midi test still performed much better than the mini anchor test across the score range of 10 to 70, where the majority of the test-takers were found. The equating bias of the midi anchor was smaller than the DTM with a few exceptions. The midi anchor even performed slightly better than the intact anchor across the entire scale range (except in the lower scale range for the chained equipercentile equating).

SG1-to-SG2 equating. The SG1-to-SG2 equating (where the two subgroups were moderately similar in ability) indicates a different pattern. As shown in Figure 3, the mini anchor now performed better than either the midi anchor or the intact anchor, either across the entire scale range (in the two linear cases) or across the majority of the scale range (in the two curvilinear cases). The mini anchor produced bias less than DTM across most of the scale range, but that was not the case for the midi anchor or the intact anchor.

More interestingly, the mini anchor even performed much better than the intact anchor, even though the latter had 15 more items (or was 75% longer) than the former and had a higher correlation with the total test. Obviously, this finding contradicts the widely held notion that a longer anchor with a higher correlation with the test to be equated results in a better equating. Between the intact anchor and the midi anchor, the former did better than the latter.

The SG1-to-SG2 equating results using data set 2 exhibited a similar pattern. As shown in Figure 4, the mini anchor did slightly better than the midi anchor, and it even did somewhat better than the intact anchor. This pattern is consistent across the equating methods: the three bias lines/curves were intertwined at the lower end of the scale, and then started to depart from each other in the middle of the scale toward the high end.

Equating methods do not seem to affect equating bias in both the total-group-to-total-group equatings and in the SG1-to-SG2 equatings. The PSE equating produced very similar or slightly larger bias than CE.

SG3-to-SG5 equating. As can be seen from Figure 5, the midi anchor performed the best for data set 1 for the SG3-to-SG5 equating. Almost across the entire scale range, the midi anchor had the smallest equating bias among the three anchor types for all the equating methods (except at the top range of the scale for the chained equipercentile method). The mini anchor performed the worst, and the intact anchor performed between the midi anchor and the mini anchor.

The results were somewhat mixed in the SG3-to-SG5 equating in data set 2. As can be seen from Figure 6, in general, the intact anchor test did the best at the lower range of the scale, whereas the midi anchor test did the best in the higher range of the scale. When comparing the midi anchor with the mini anchor, the former did better than the latter either across the entire score range (in the two linear equatings) or across most of the score range (in the two curvilinear equatings).

SG4-to-SG5 equating. Figure 7 exhibits the results of SG4-to-SG5 equating from data set 1. The pattern was quite similar to the pattern we observed in the total-group-to-total-group equating: the three types of anchors performed in a similar fashion, with the intact anchor performing slightly better than the other two. Between the mini anchor and the midi anchor, the midi anchor did slightly better than the mini anchor.

The bias results of SG4-to-SG5 equating from data set 2 are shown in Figure 8. The intact anchor produced the most accurate results, while the mini anchor yielded the least accurate results. The midi anchor yielded smaller bias than the mini anchor virtually across the entire scale range, and across all of the four equating methods.

Equating method has impact on the equating bias in both SG3-to-SG5 equatings and in SG4-to-SG5 equatings. The CE is less biased than the PSE.

Among the four pairs of equating samples, the total-group-to-total-group equating produced the smallest conditional equating bias, whereas the SG4-to-SG5 equating yielded the largest conditional equating bias.

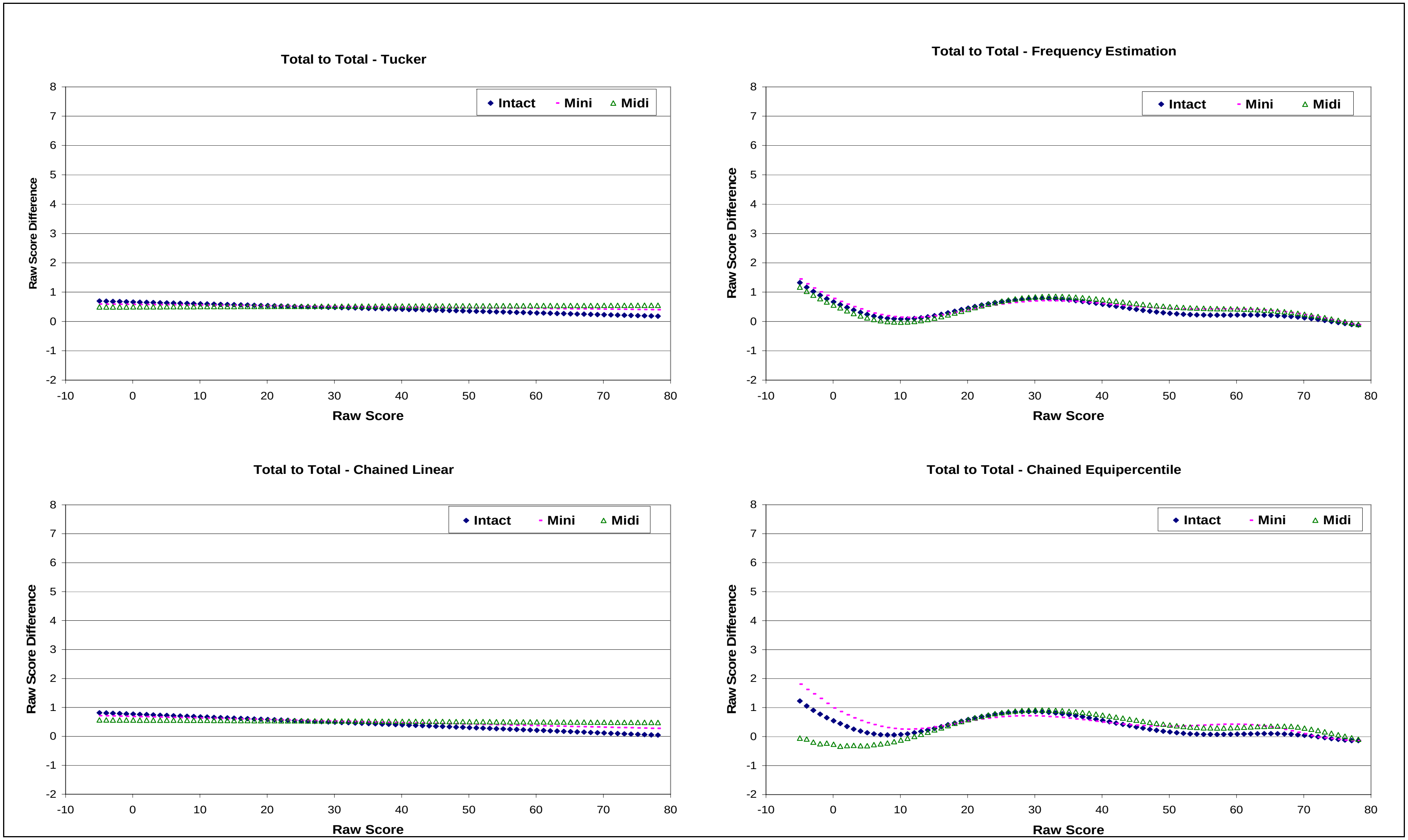


Figure 1. **Bias in the total-group-to-total-group equating: *X1* to *Y1*.**

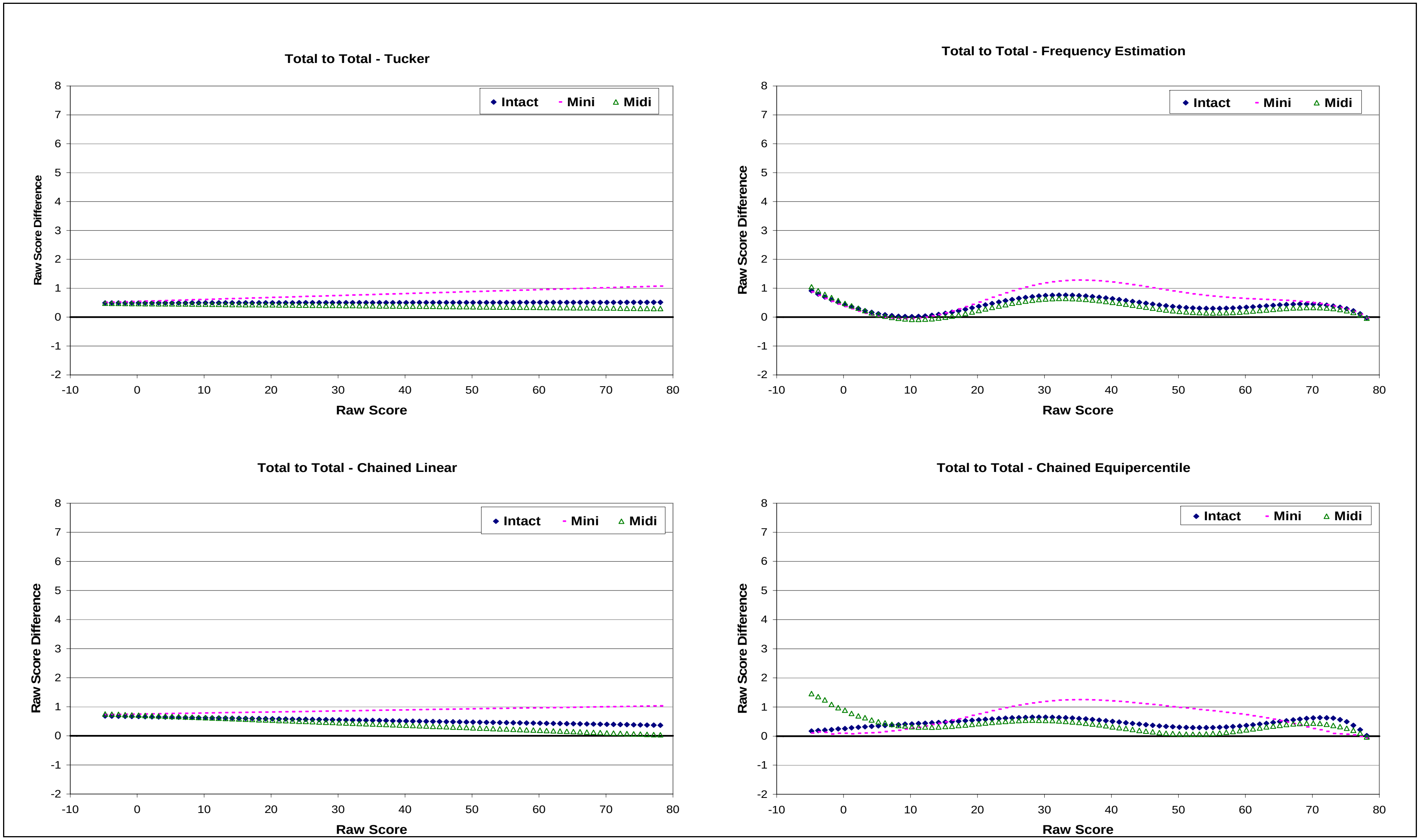
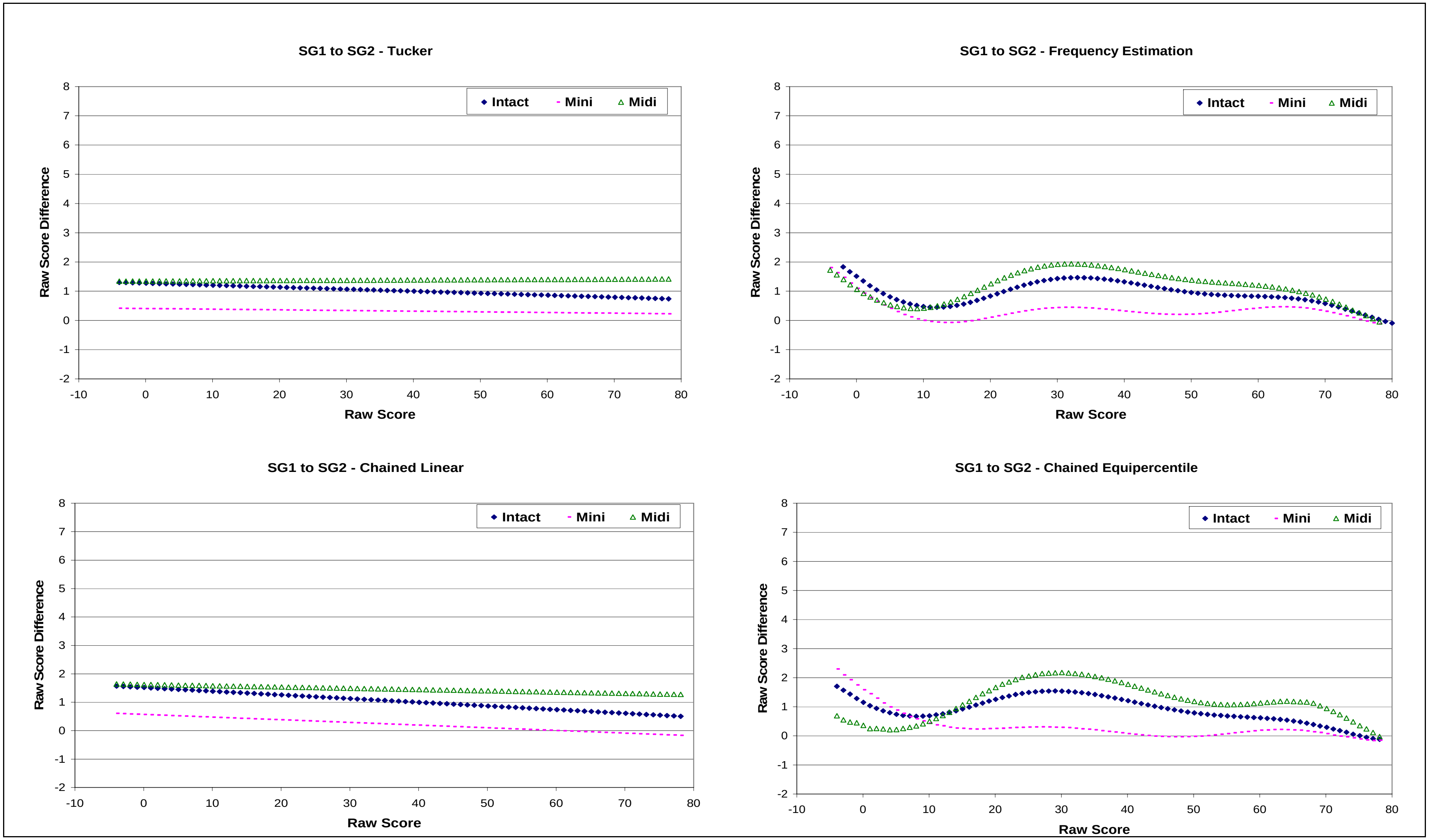


Figure 2. **Bias in the total-group-to-total-group equating: *X2* to *Y2*.**


SG1 to SG2 - Tucker

SG1 to SG2 - Frequency Estimation

SG1 to SG2 - Chained Linear

SG1 to SG2 - Chained Equipercentile

Figure 3. **Bias in the SG1-to-SG2 equating: *X1* to *Y1*.**

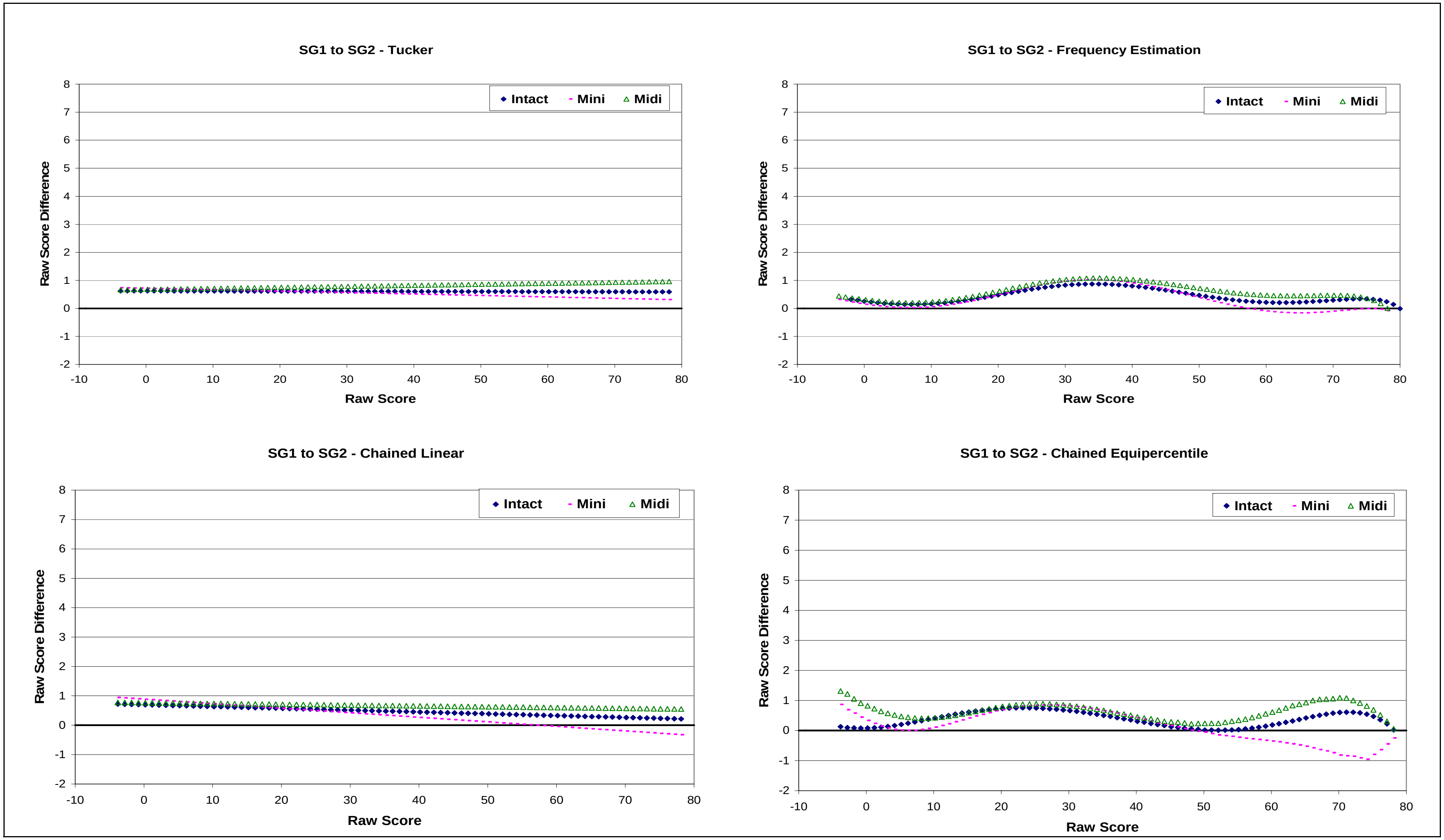


Figure 4. **Bias in the SG1-to-SG2 equating: *X2* to *Y2*.**

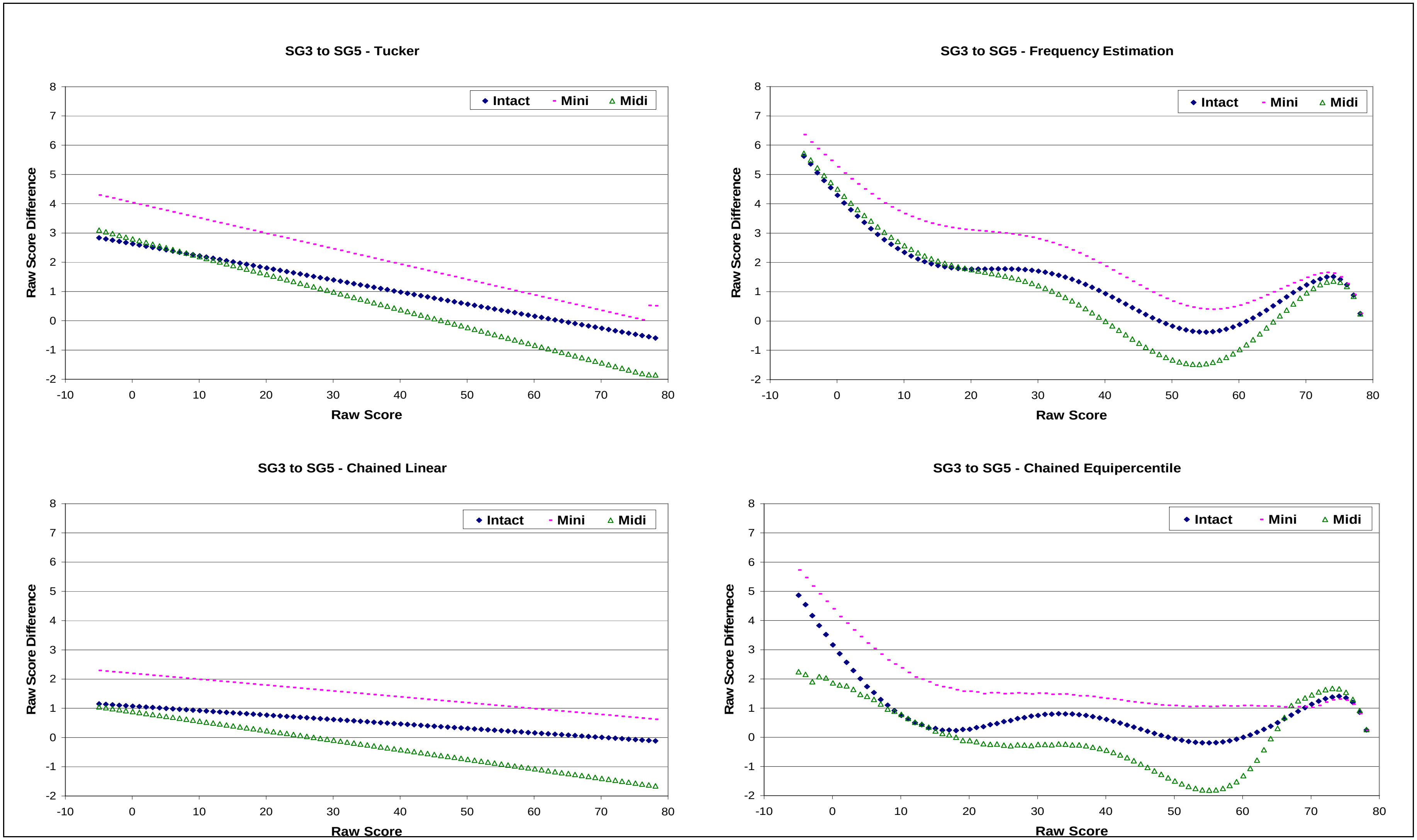


Figure 5. **Bias in the SG3-to-SG5 equating: *X1* to *Y1*.**

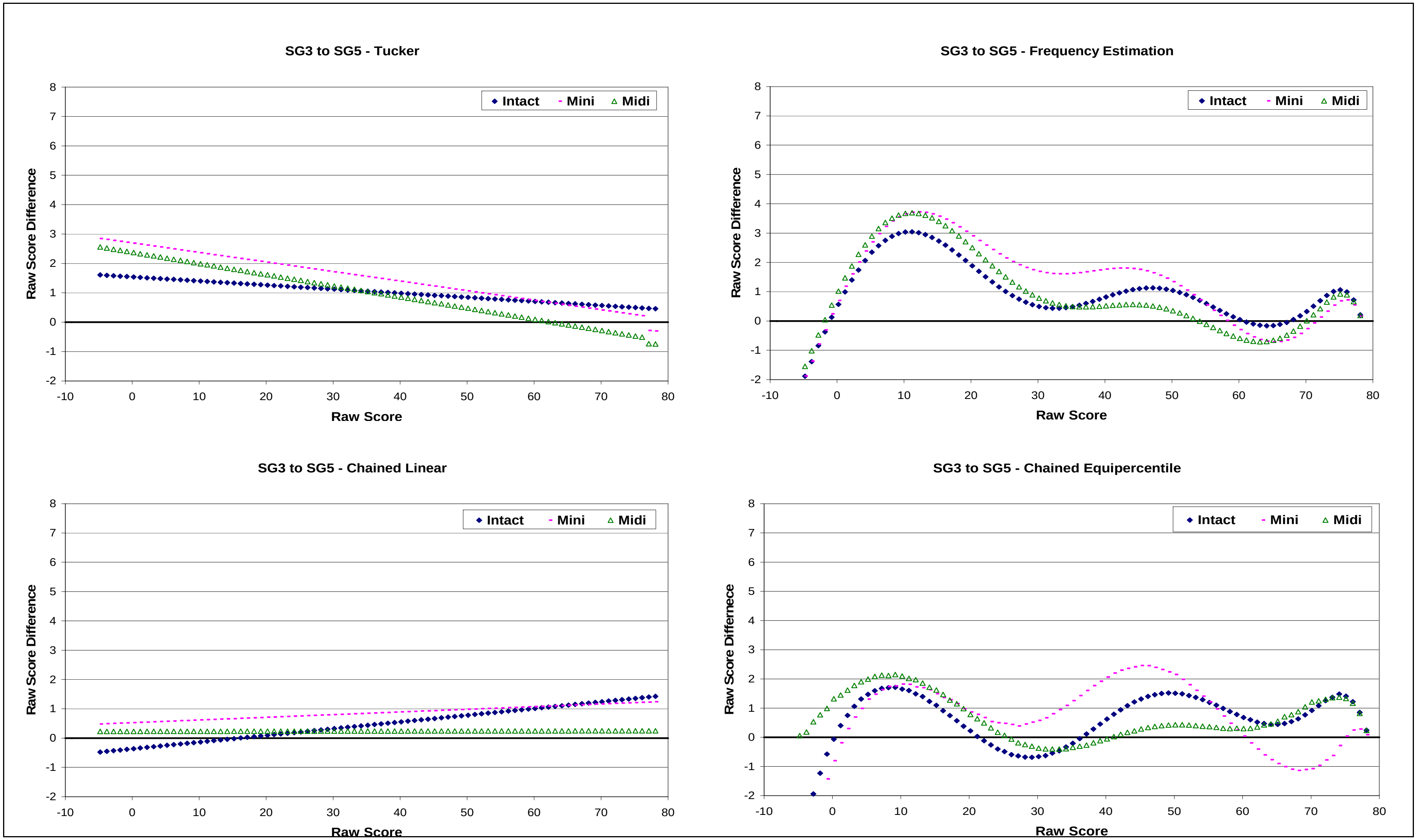


Figure 6. **Bias in the SG3-to-SG5 equating: *X2* to *Y2*.**

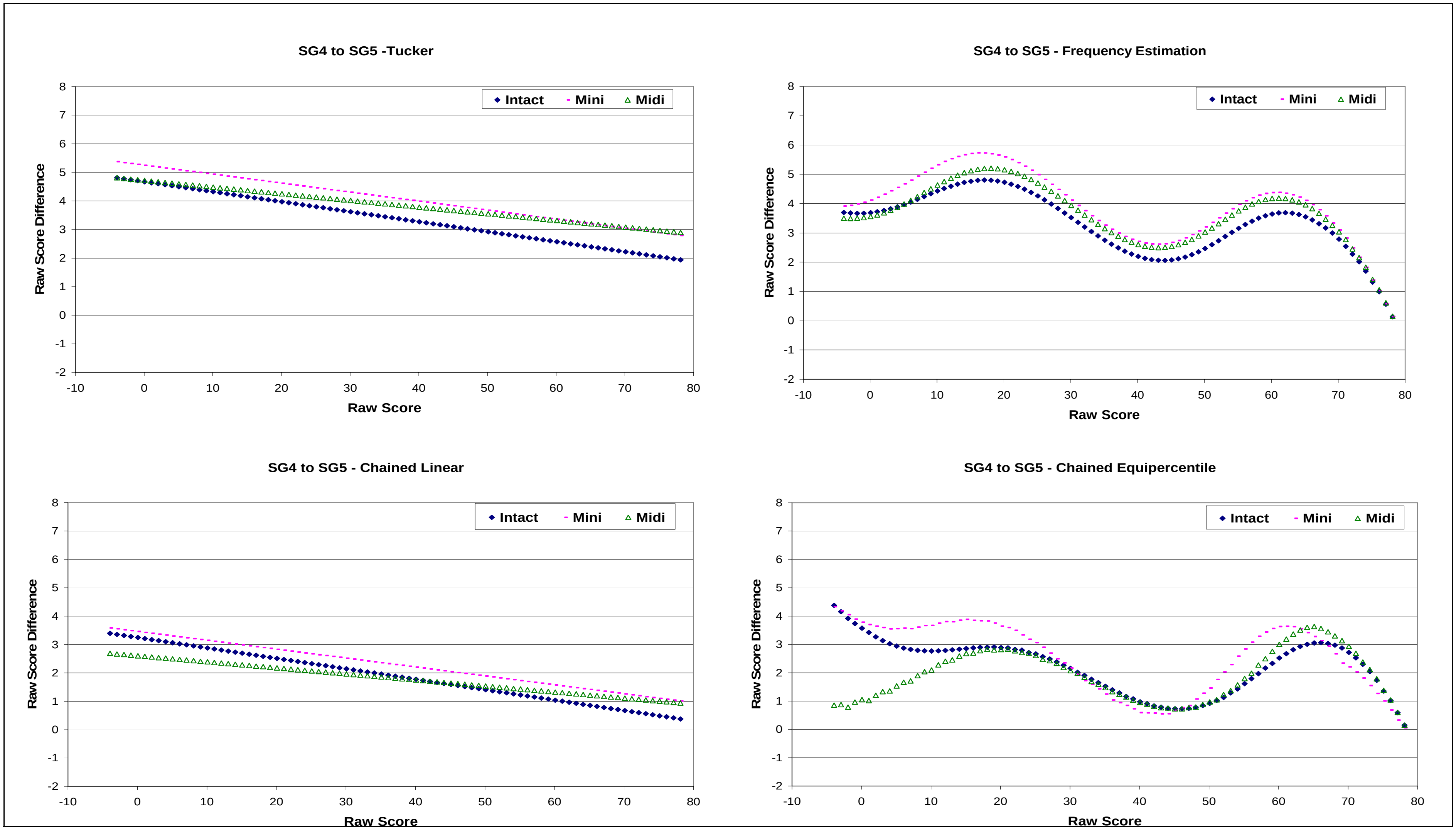


Figure 7. **Bias in the SG4-to-SG5 equating: *X1* to *Y1*.**

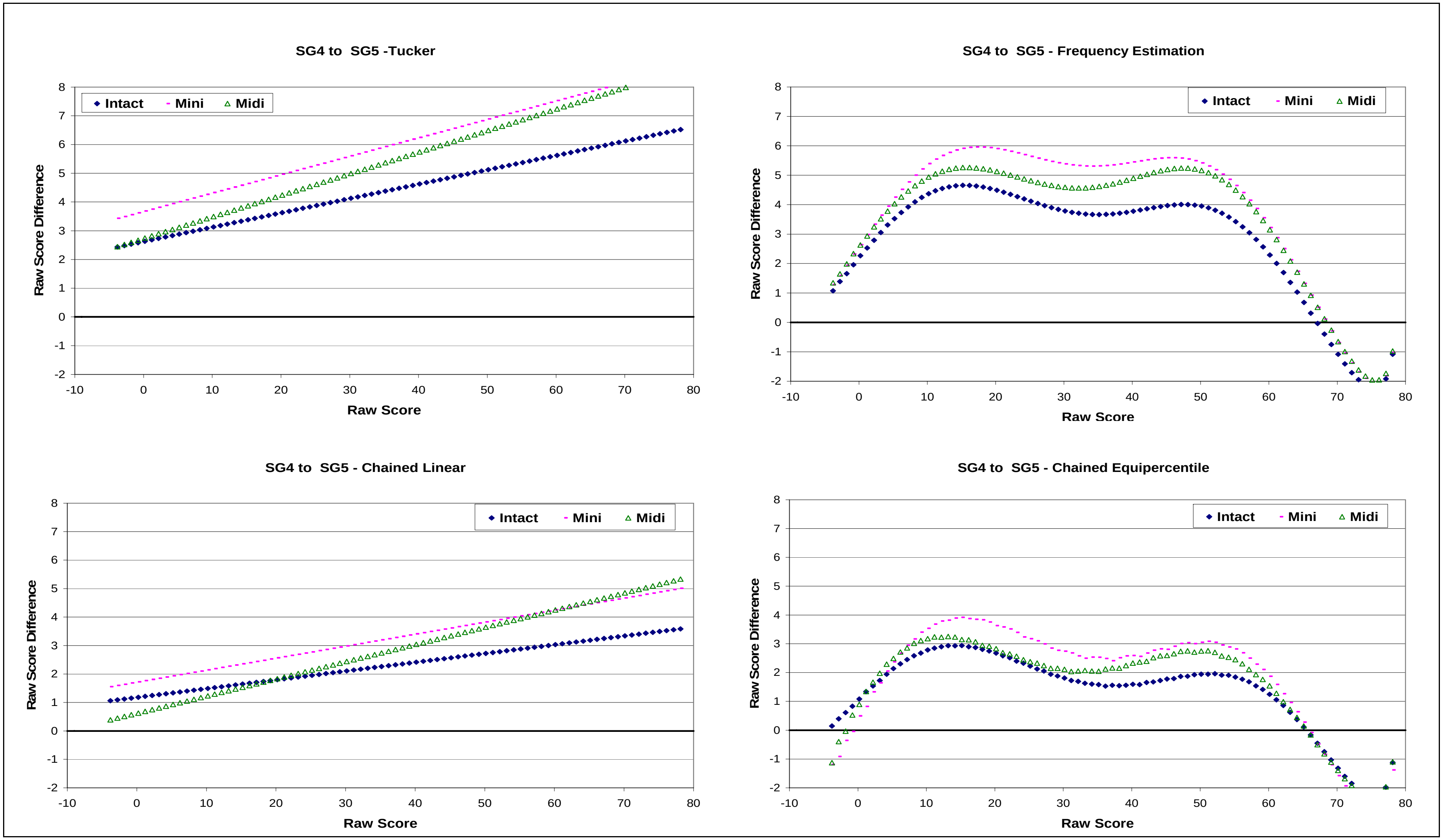


Figure 8. **Bias in the SG4-to-SG5 equating: *X2* to *Y2*.**

Standard Errors of Equating (SEE)

The SEEs are presented in Figures 9 to 16. All the figures show a consistent pattern: the intact anchor produced the smallest SEE. The midi anchor and the mini anchor showed very similar results: both produced slightly larger SEE than the intact anchor. Among the four pairs of equating samples, the total-group-to-total-group equatings produced the smallest SEE, while the SG3-to-SG5 and SG4-to-SG5 equatings produced larger equating errors. Across the equating methods, the chained equipercentile equatings tended to produce the largest SEE.

Root Mean Squared Error (RMSE)

Figures 17 to 24 present the RMSE curves. Tables 10 and 11 summarize the weighted absolute bias, the weighted SEE, and the weighted RMSE for data sets 1 and 2, respectively.

Total-group-to-total-group equating. In the total-group-to-total-group equating for data set 1, as shown Figure 17, the three RMSE curves are very similar to each other, with the intact anchor doing slightly better at the higher end of the score range. Table 10 shows that the intact anchor had the smallest weighted absolute equating bias, followed by the mini anchor and then the midi anchor. The intact anchor also had the smallest SEE among the three anchors, although the differences were very small. Correspondingly, the intact anchor produced the smallest average RMSE. Between the mini anchor and the midi anchor, the mini anchor did slightly better.

Table 11 and Figure 18 summarize and present the total-group-to-total-group results for data set 2. The difference from data set 1 is that the midi anchor yielded the smallest RMSE curve across all of the four equating methods. The midi anchor also produced the smallest average bias and RMSE, even though the intact anchor had the smallest SEE. It is clear that the mini anchor did the worst.

SG1-to-SG2 equating. In the SG1-to-SG2 equating for data set 1 (Table 10 and Figure 19) and data set 2 (Table 11 and Figure 20), the mini anchor produced the smallest equating bias and RMSE, even smaller than those of the intact anchor. The intact anchor did worse than the mini anchor, but it did better than the midi anchor.

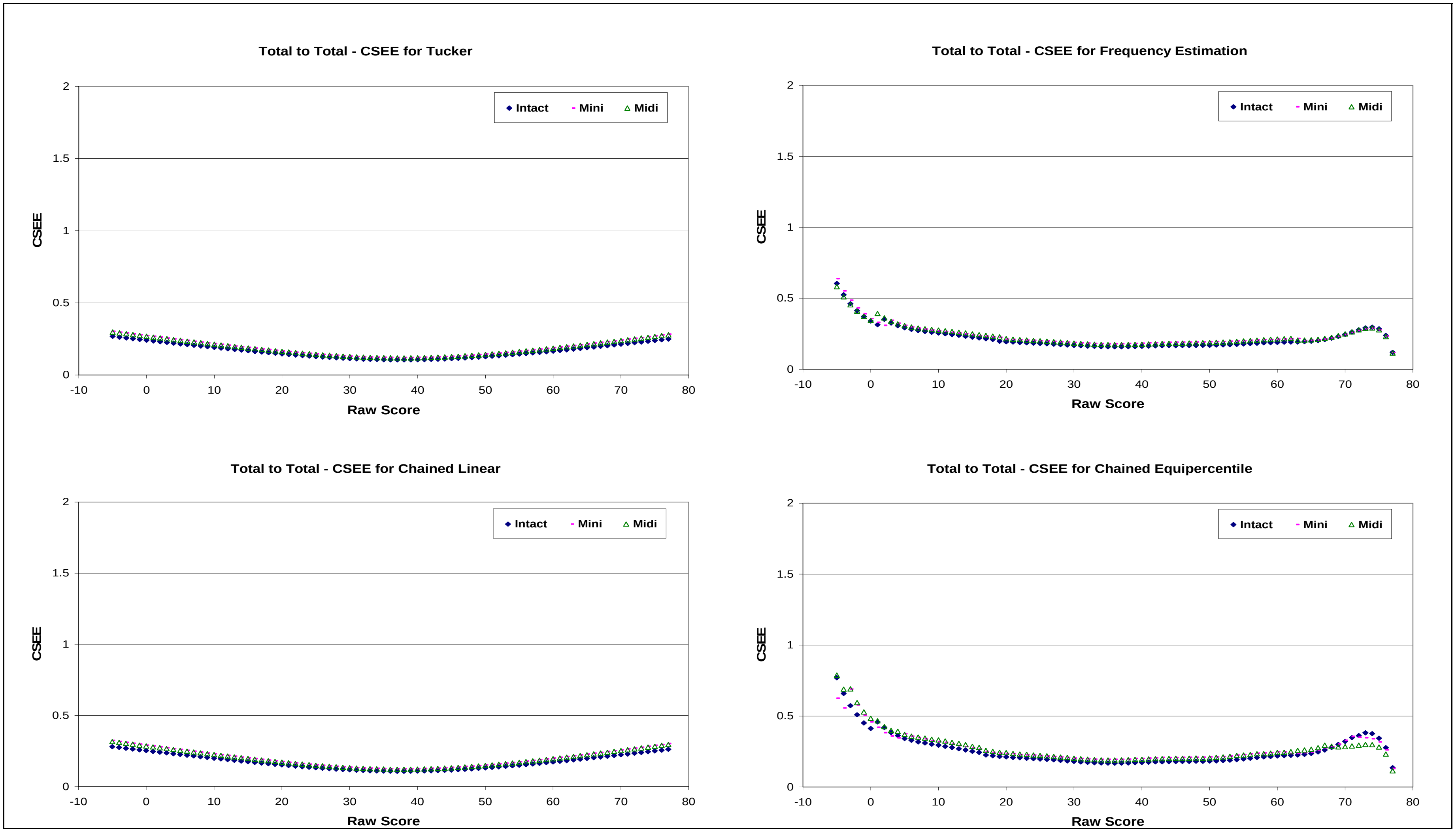


Figure 9. **Standard errors of equating (SEE) in the total-group-to-total-group equating: *X1* to *Y1*.**

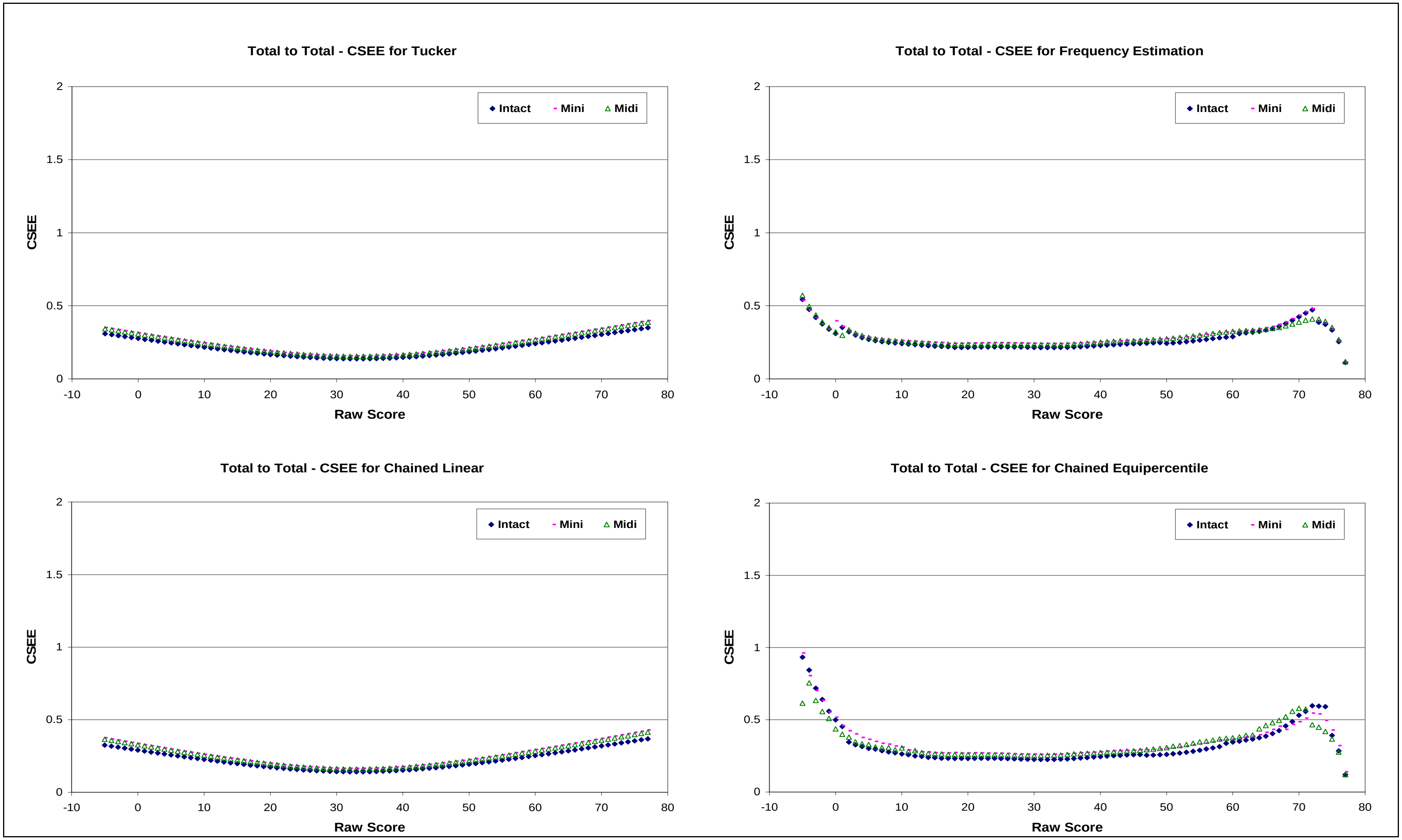


Figure 10. **Standard errors of equating (SEE) in the total-group-to-total-group equating: *X2* to *Y2*.**

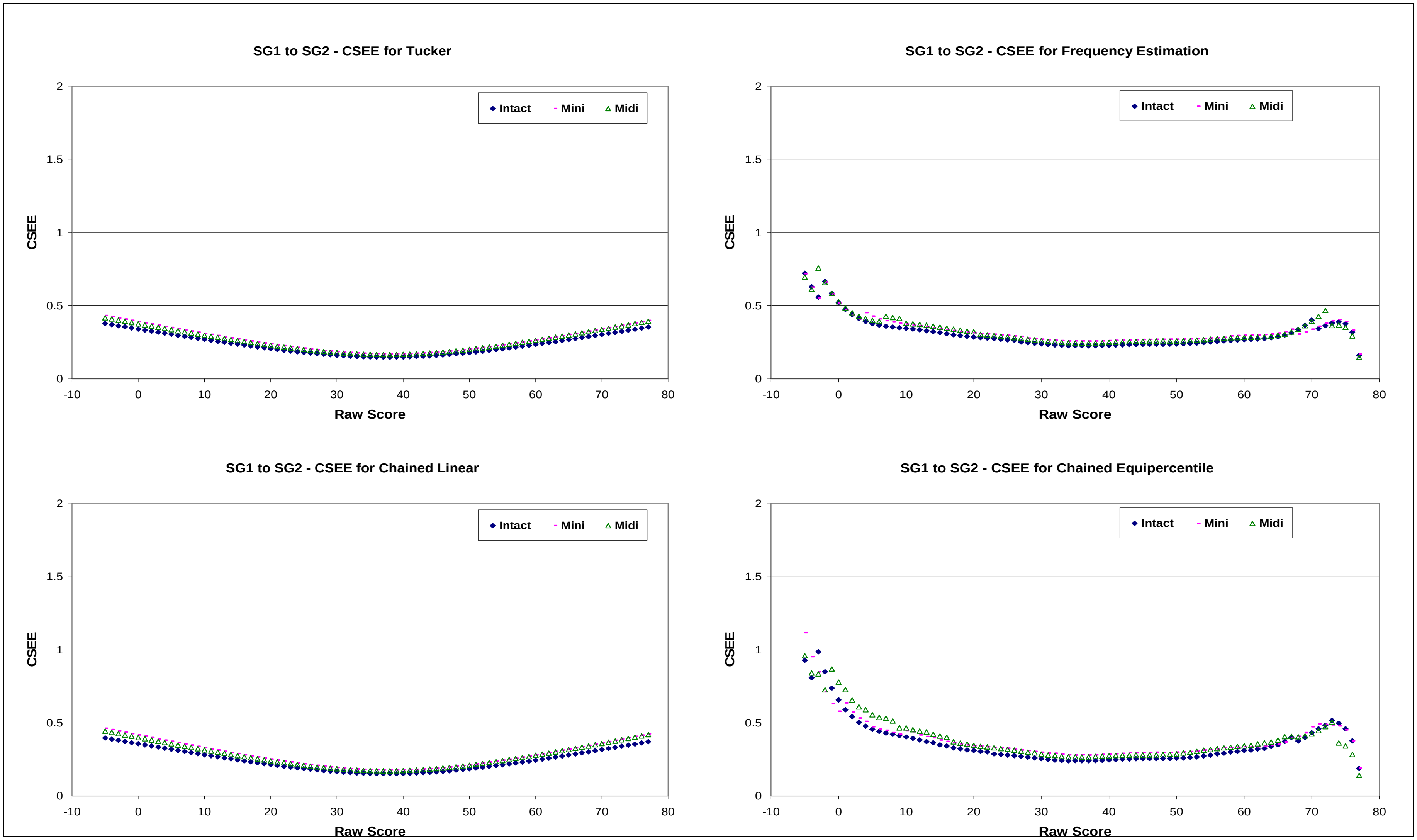


Figure 11. **Standard errors of equating (SEE) in the SG1-to-SG2 equating: *X1* to *Y1*.**

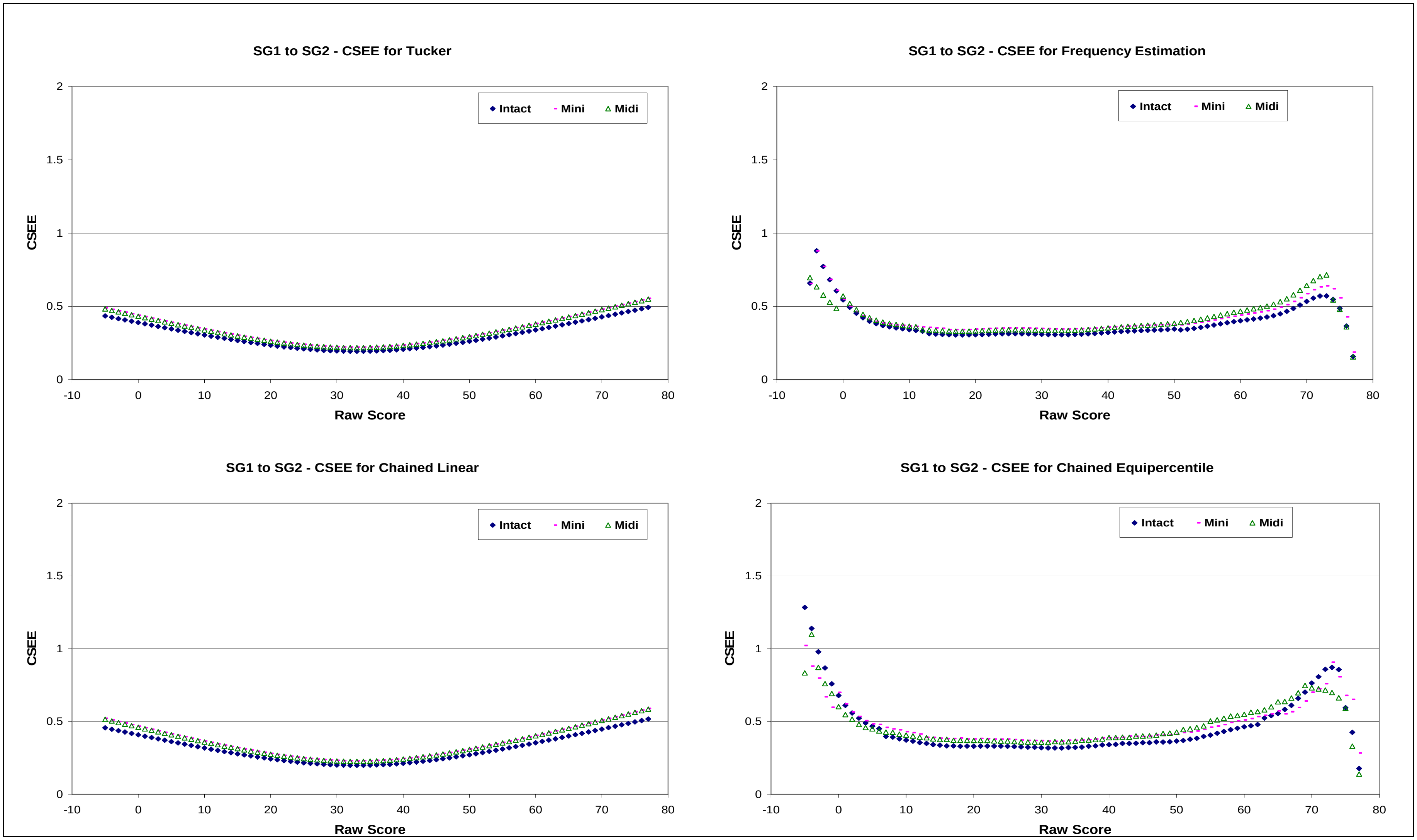


Figure 12. **Standard errors of equating (SEE) in the SG1-to-SG2 equating: *X2* to *Y2*.**

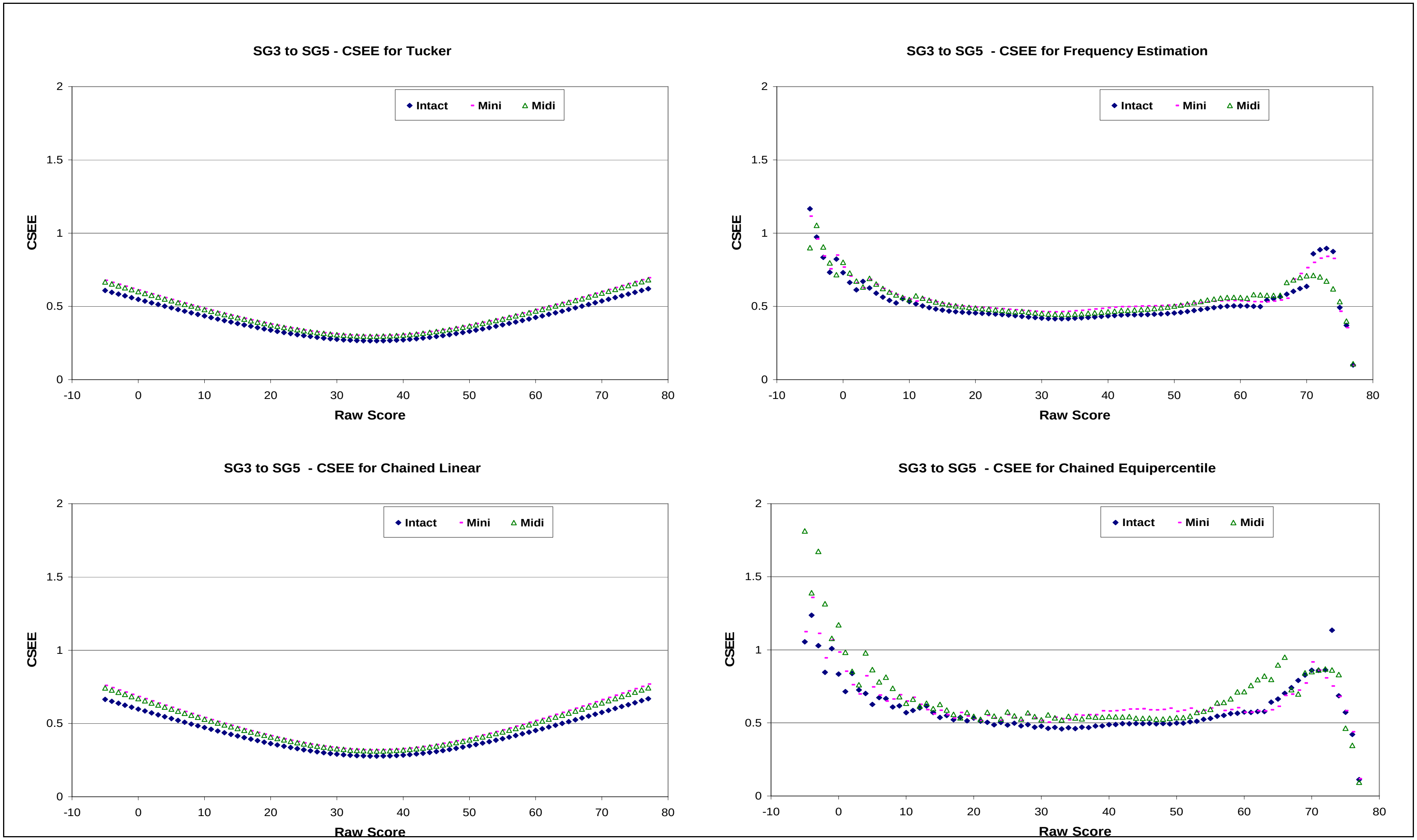


Figure 13. **Standard errors of equating (SEE) in the SG3-to-SG5 equating: *X1* to *Y1*.**

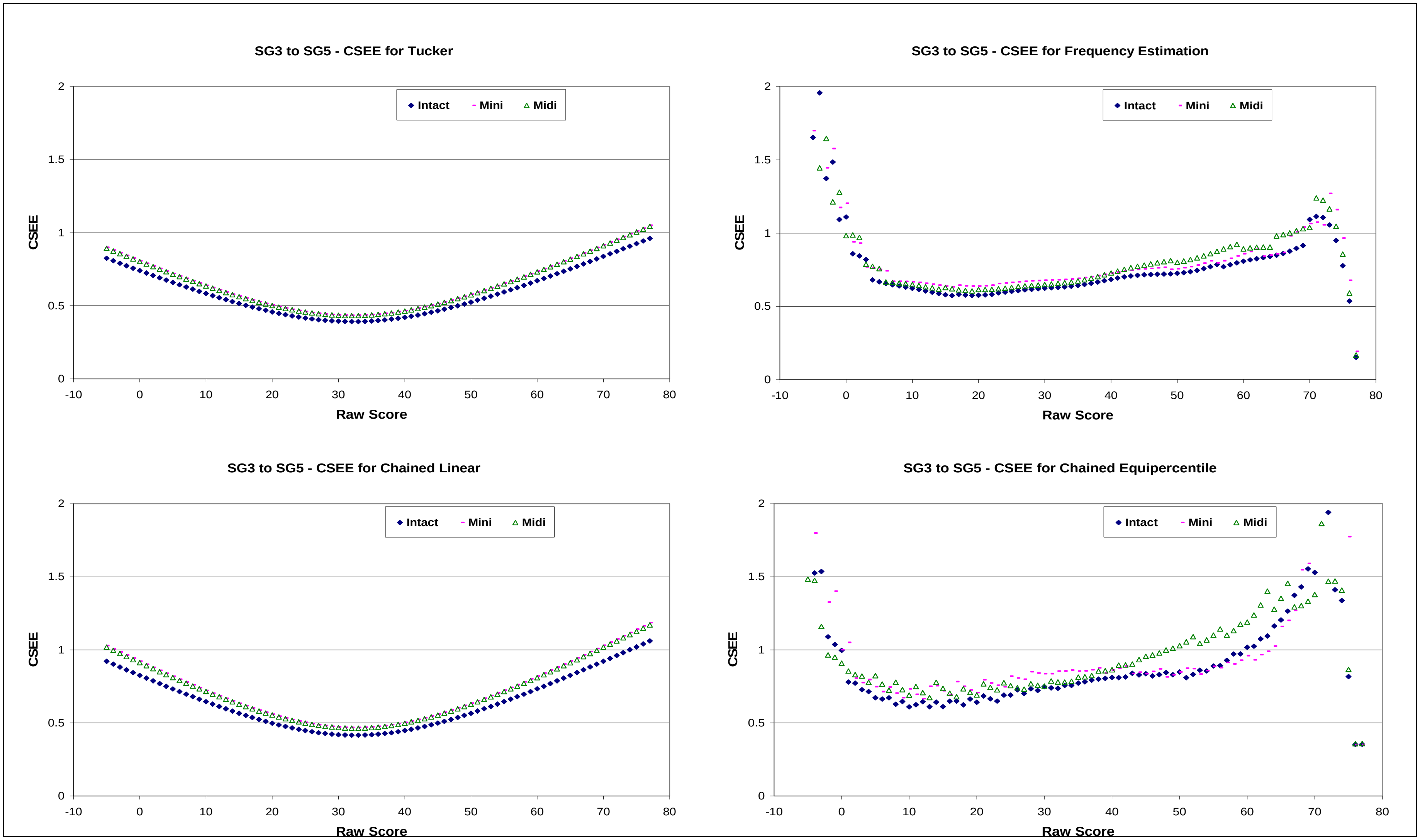


Figure 14. **Standard errors of equating (SEE) in the SG3-to-SG5 equating: *X2* to *Y2*.**

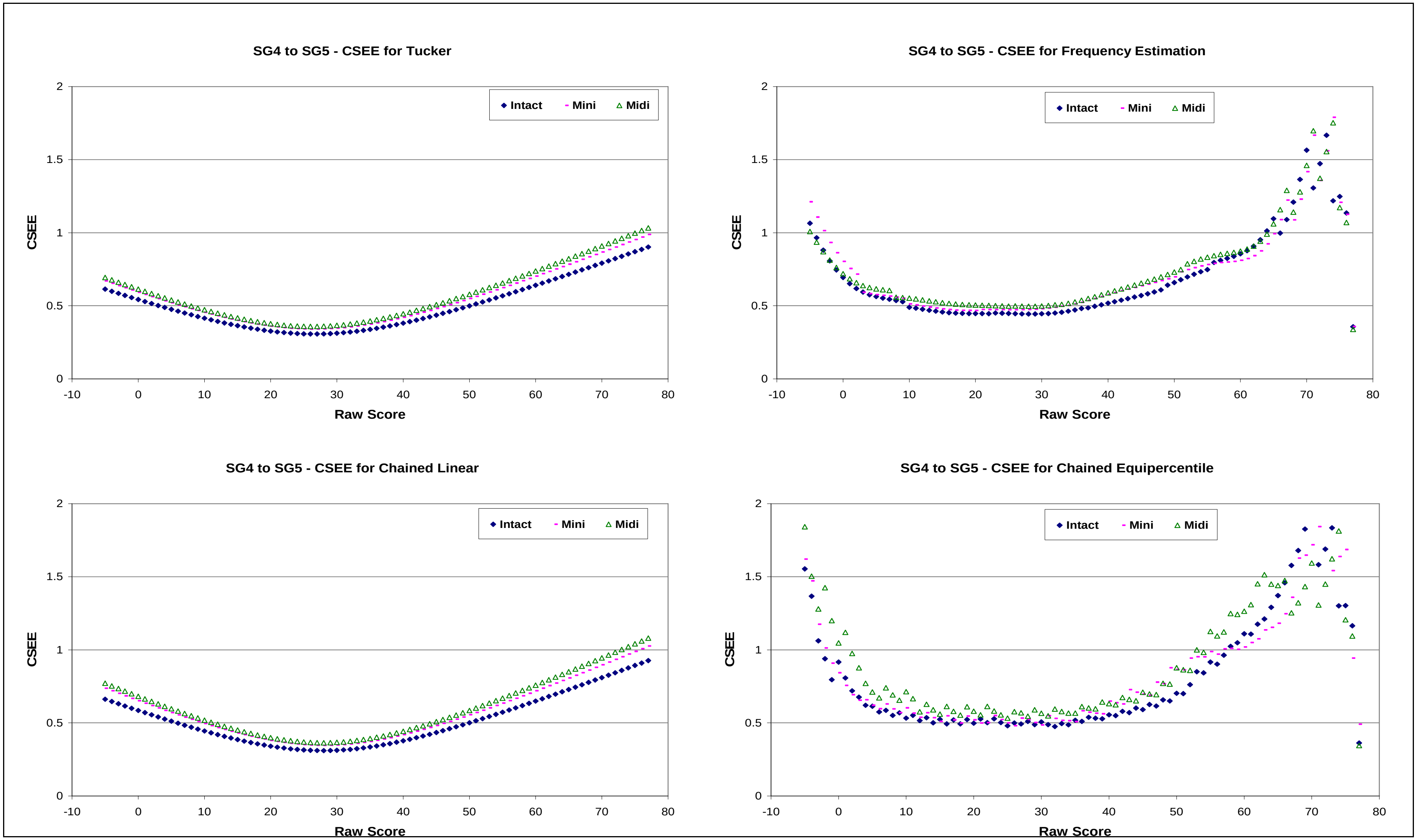


Figure 15. **Standard errors of equating (SEE) in the SG4-to-SG5 equating: *X1* to *Y1*.**

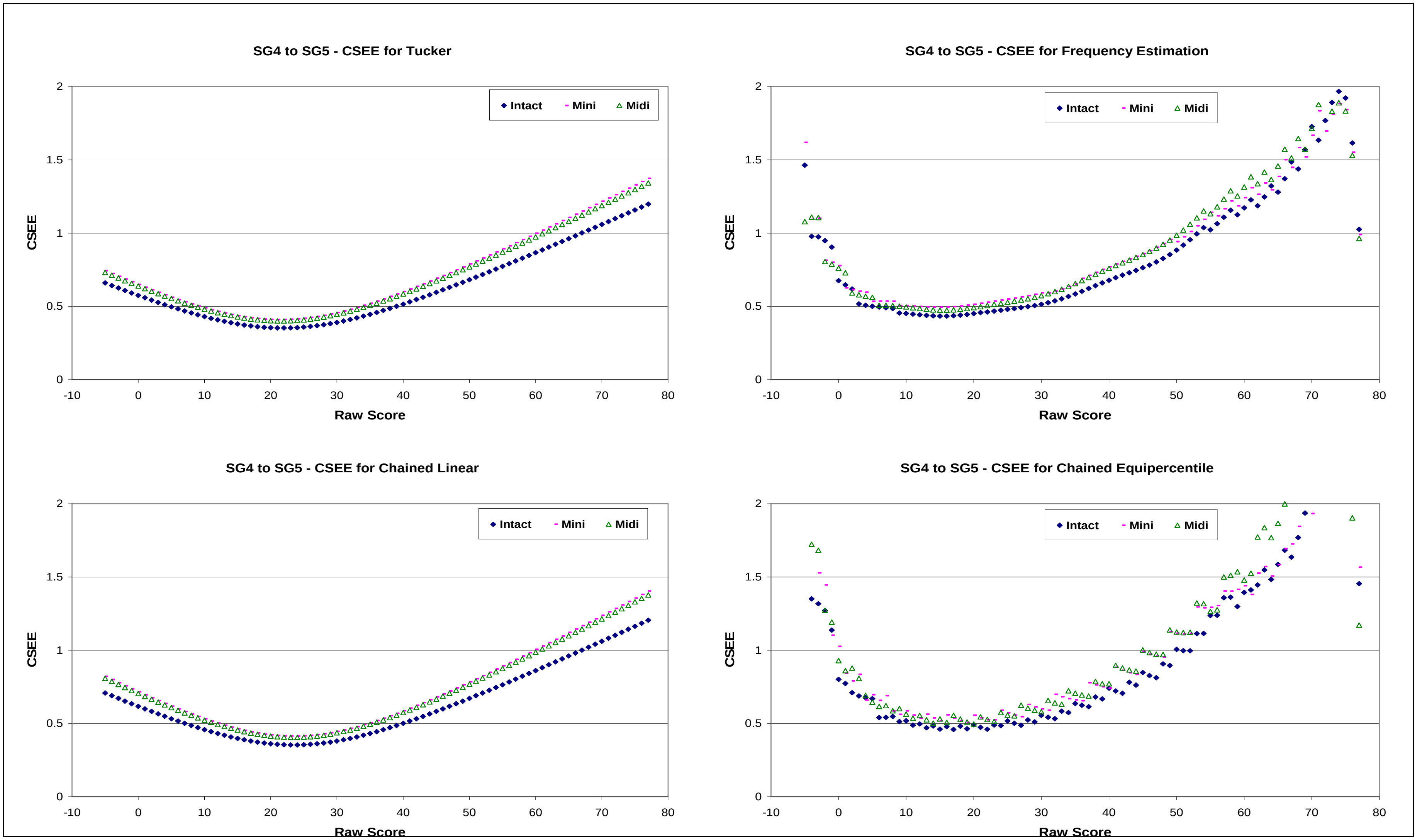


Figure 16. Standard errors in equating (SEE) in the SG4-to-SG5 equating: *X2* to *Y2*.

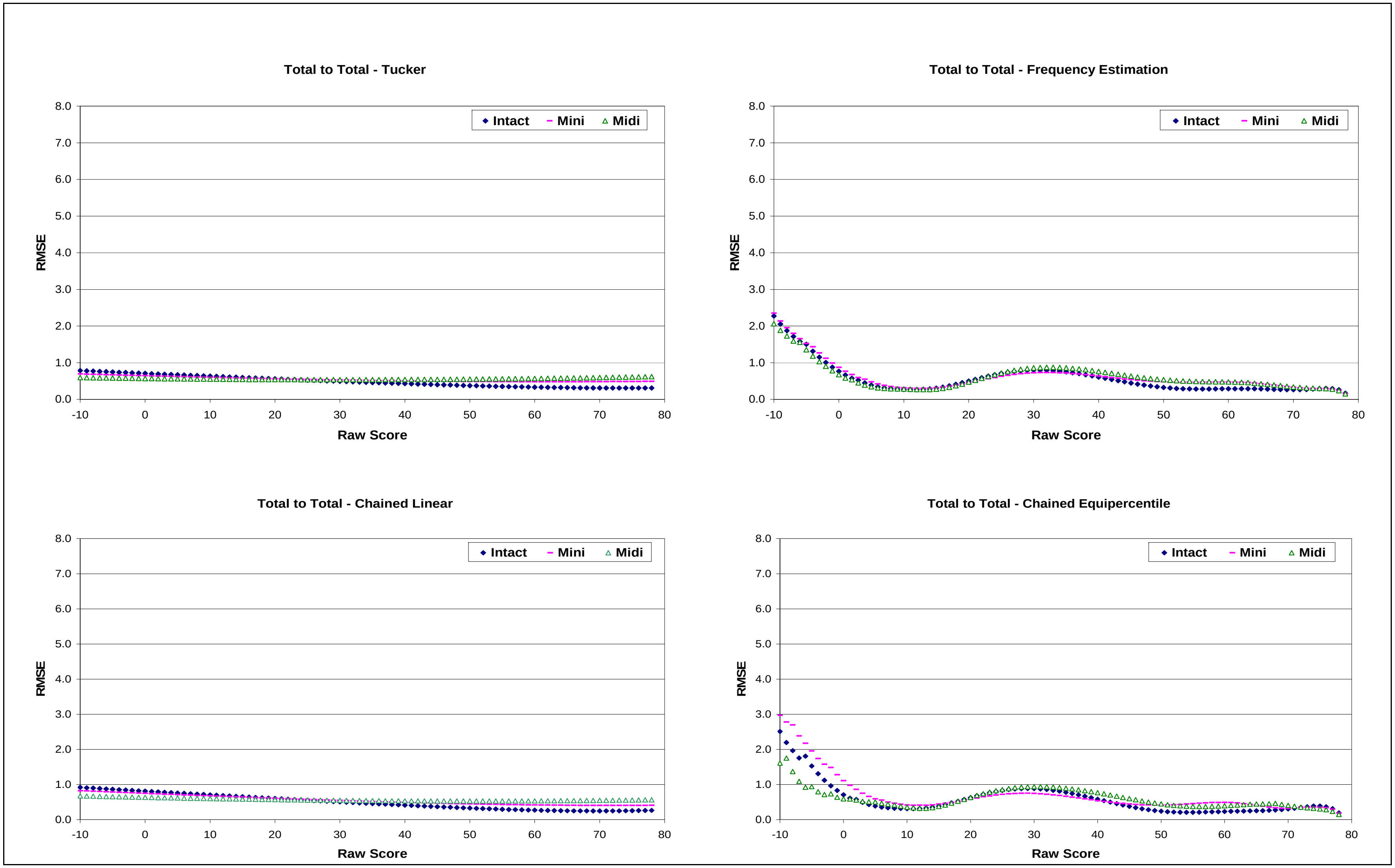


Figure 17. **Root mean squared error (RMSE) in the total-group-to-total-group equating: *X1* to *Y1*.**

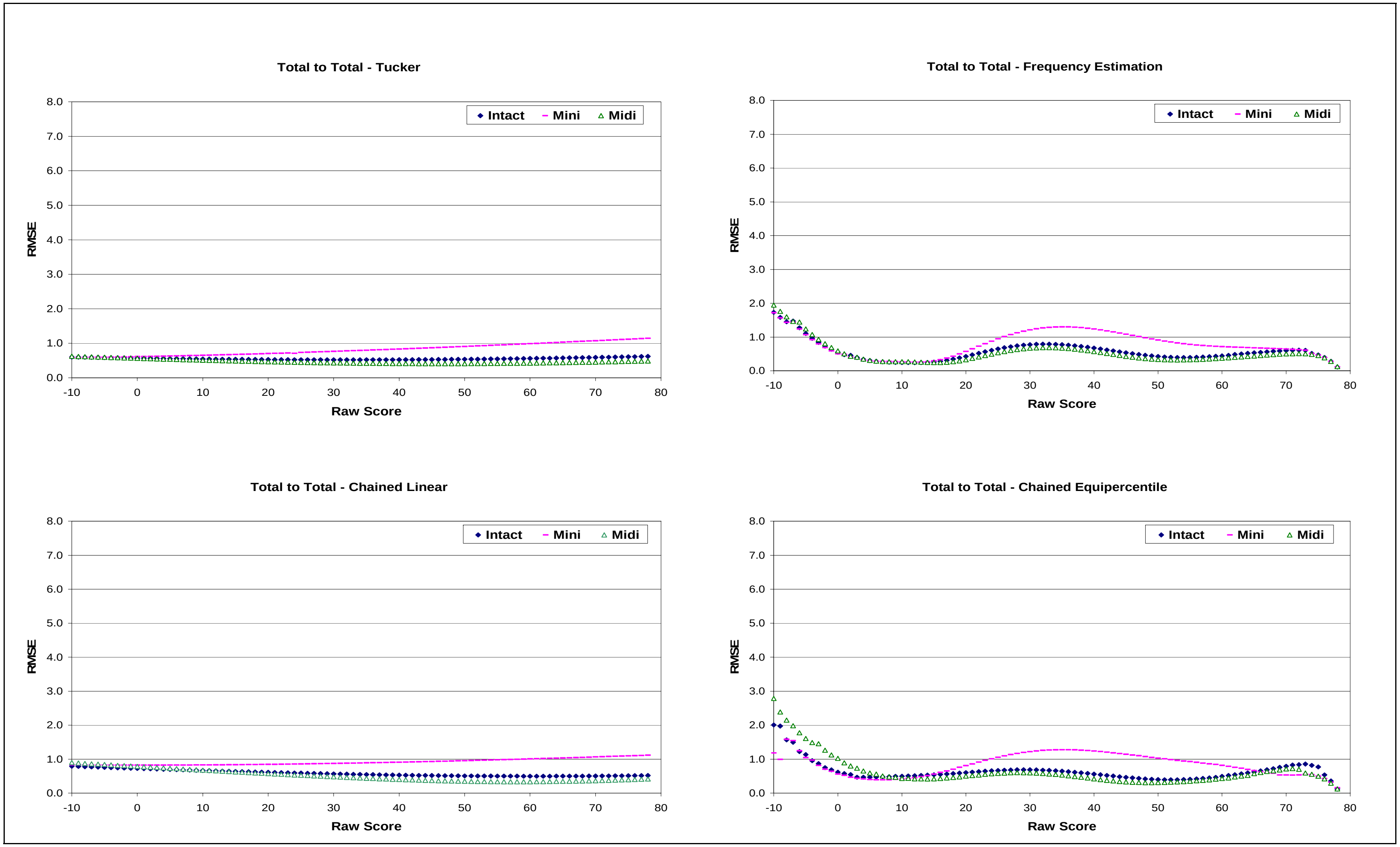


Figure 18. **Root mean squared error (RMSE) in the total-group-to-total-group equating: *X2* to *Y2*.**

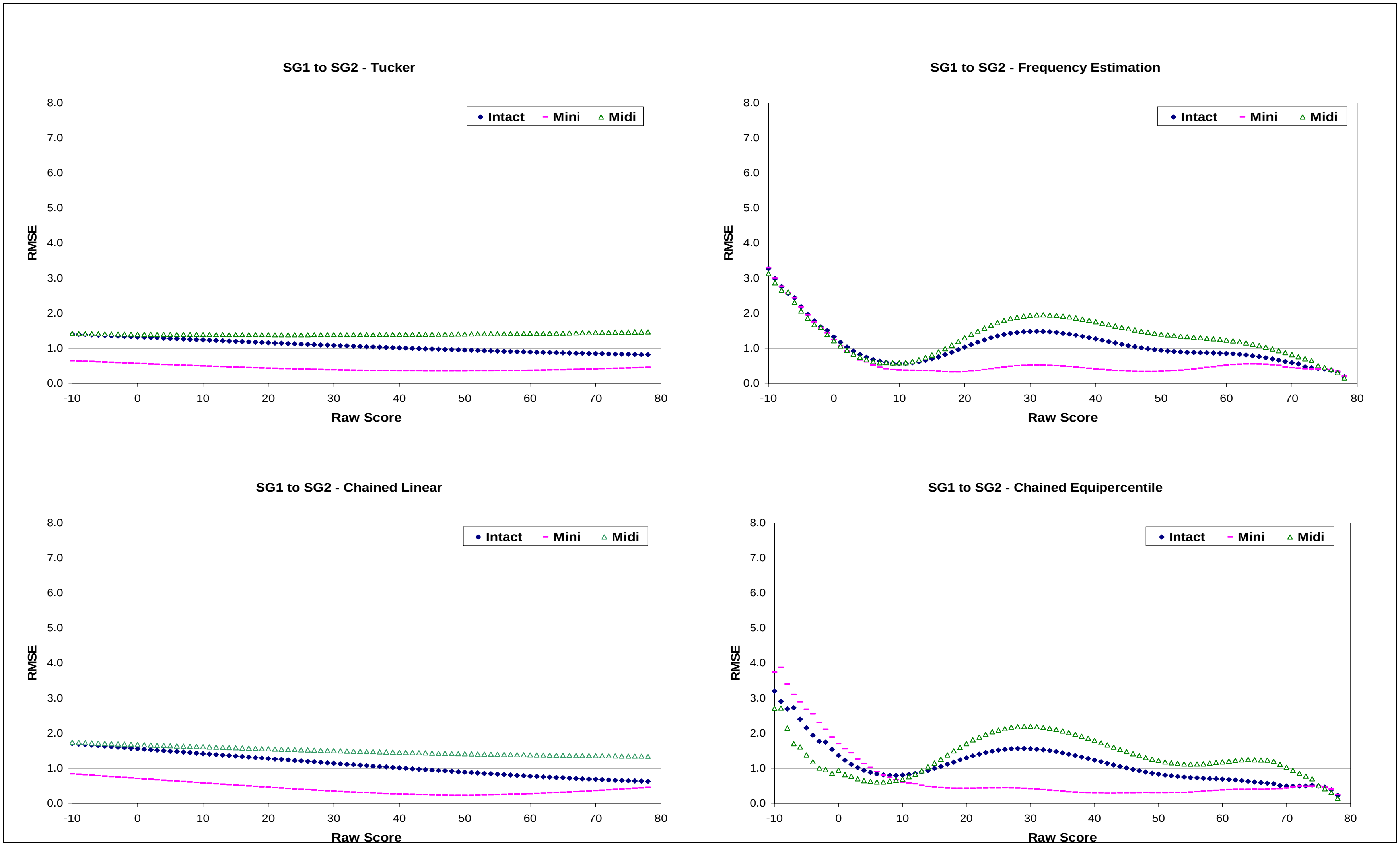
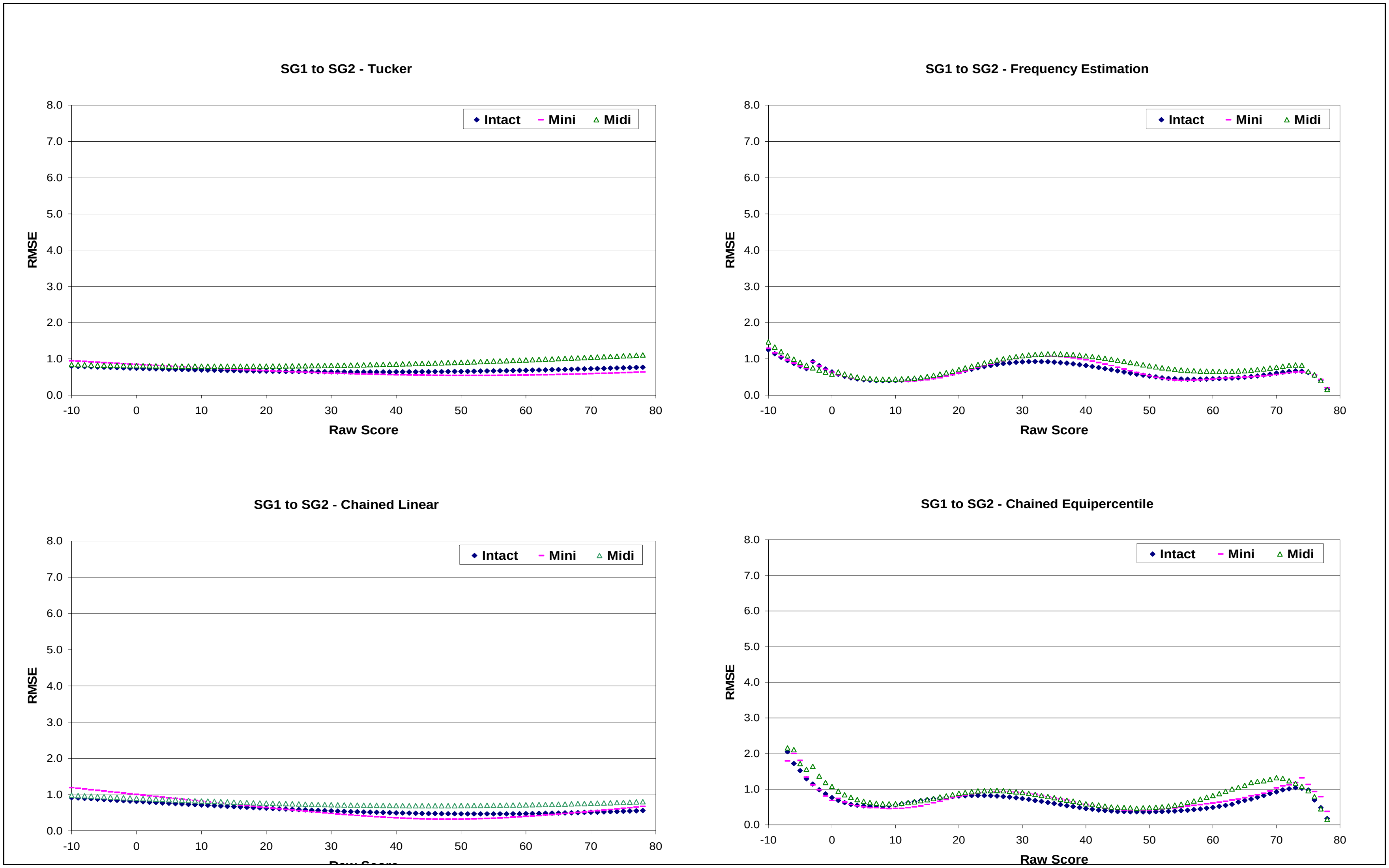


Figure 19. **Root mean squared error (RMSE) in the SG1-to-SG2 equating: *X1* to *Y1*.**



SG1 to SG2 - Tucker

SG1 to SG2 - Frequency Estimation

SG1 to SG2 - Chained Linear

SG1 to SG2 - Chained Equipercentile

Figure 20. **Root mean squared error (RMSE) in the SG1-to-SG2 equating: *X2* to *Y2*.**

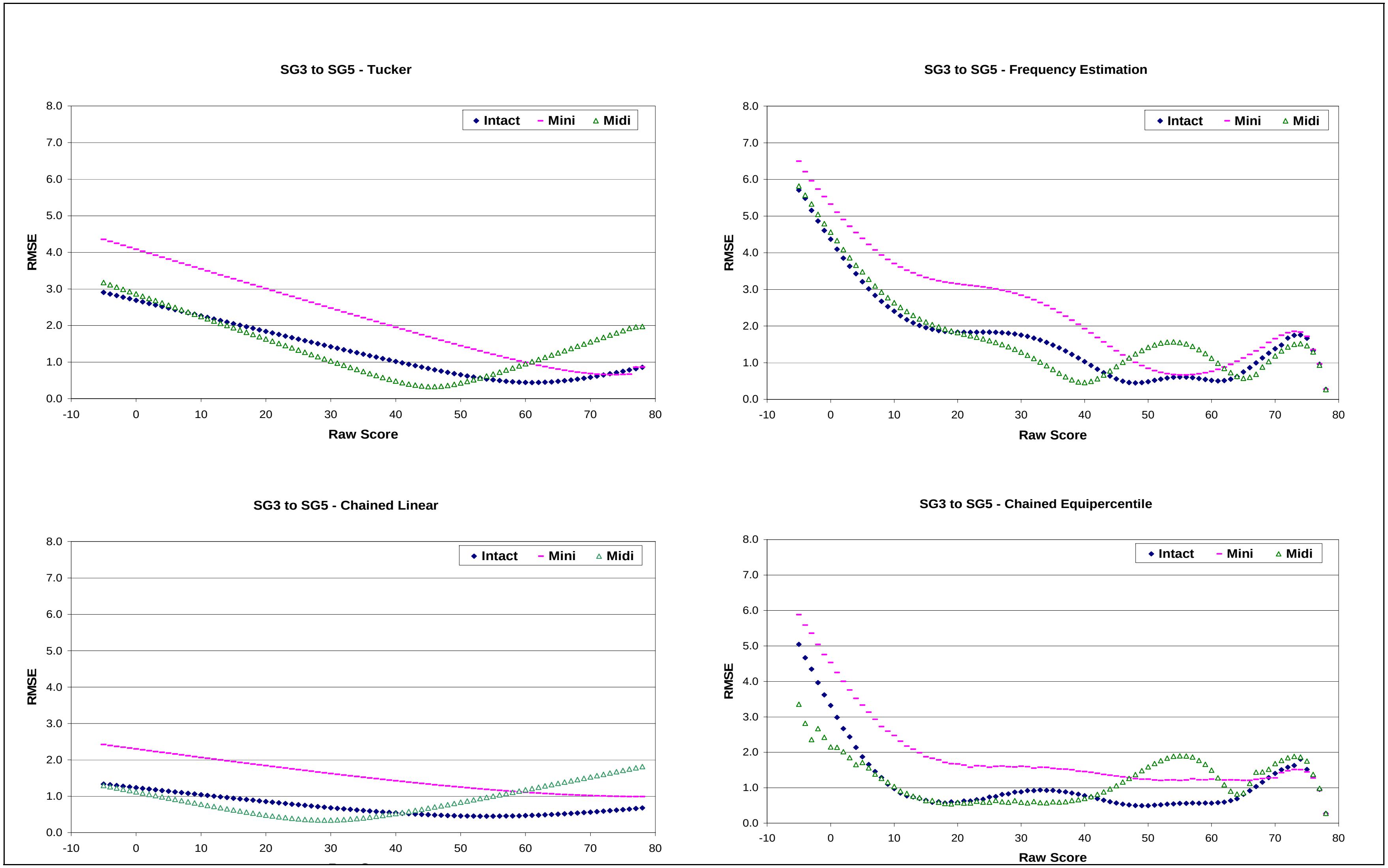


Figure 21. **Root mean squared error (RMSE) in the SG3-to-SG5 equating: *X1* to *Y1***

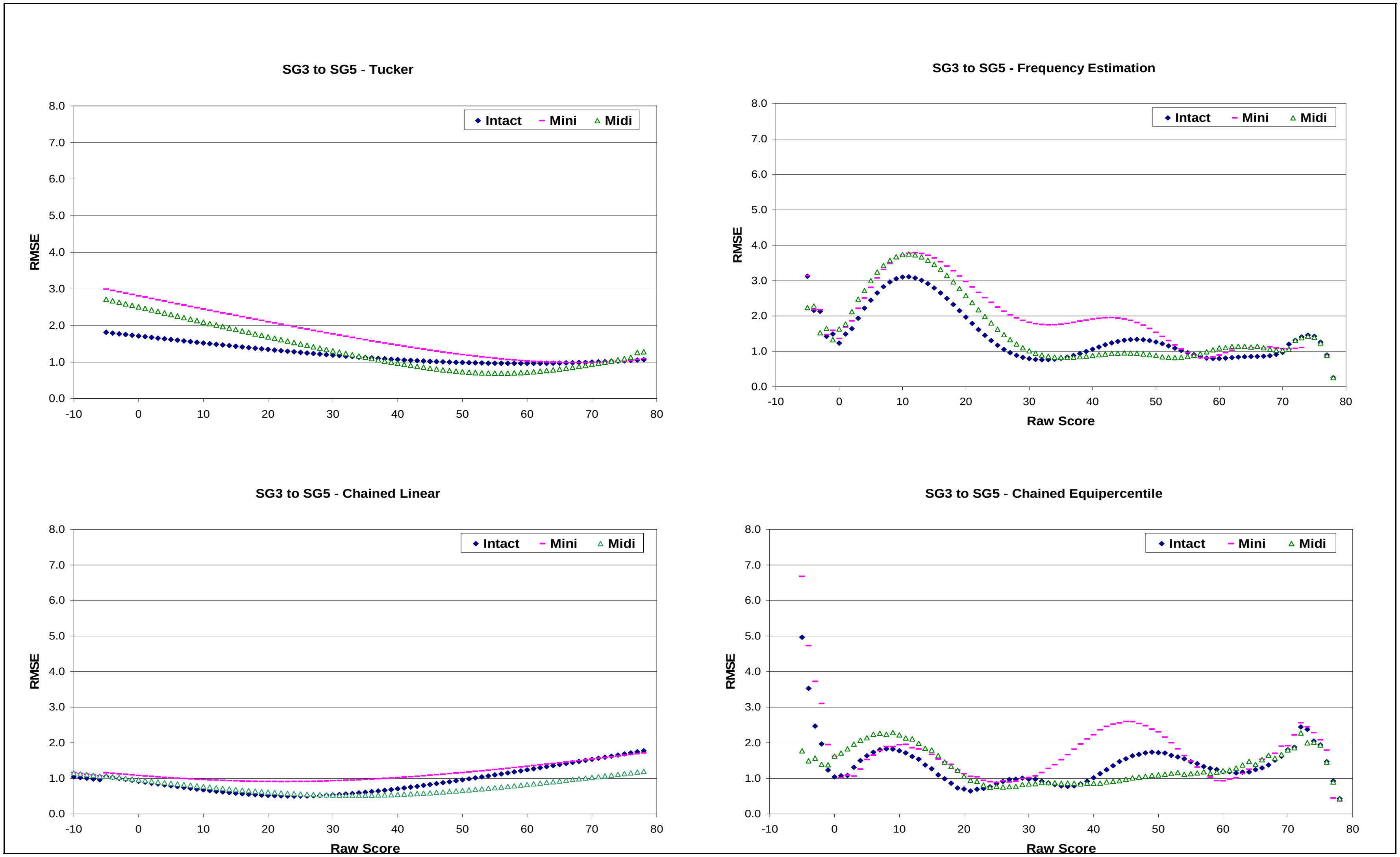


Figure 22. **Root mean squared error (RMSE) in the SG3-to-SG5 equating: *X2* to *Y2*.**

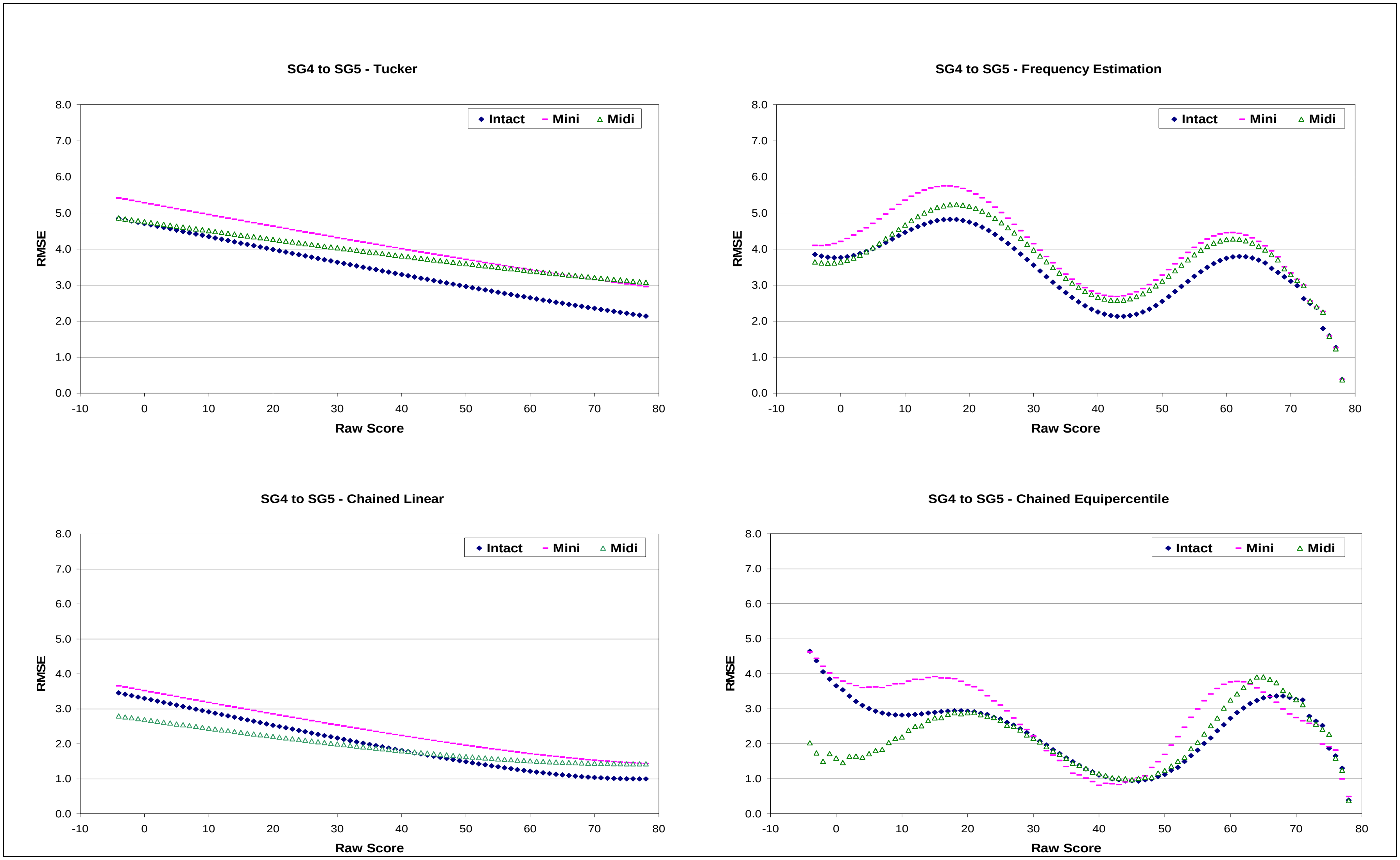


Figure 23. **Root mean squared error (RMSE) in the SG4-to-SG5 equating: *X1* to *Y1*.**

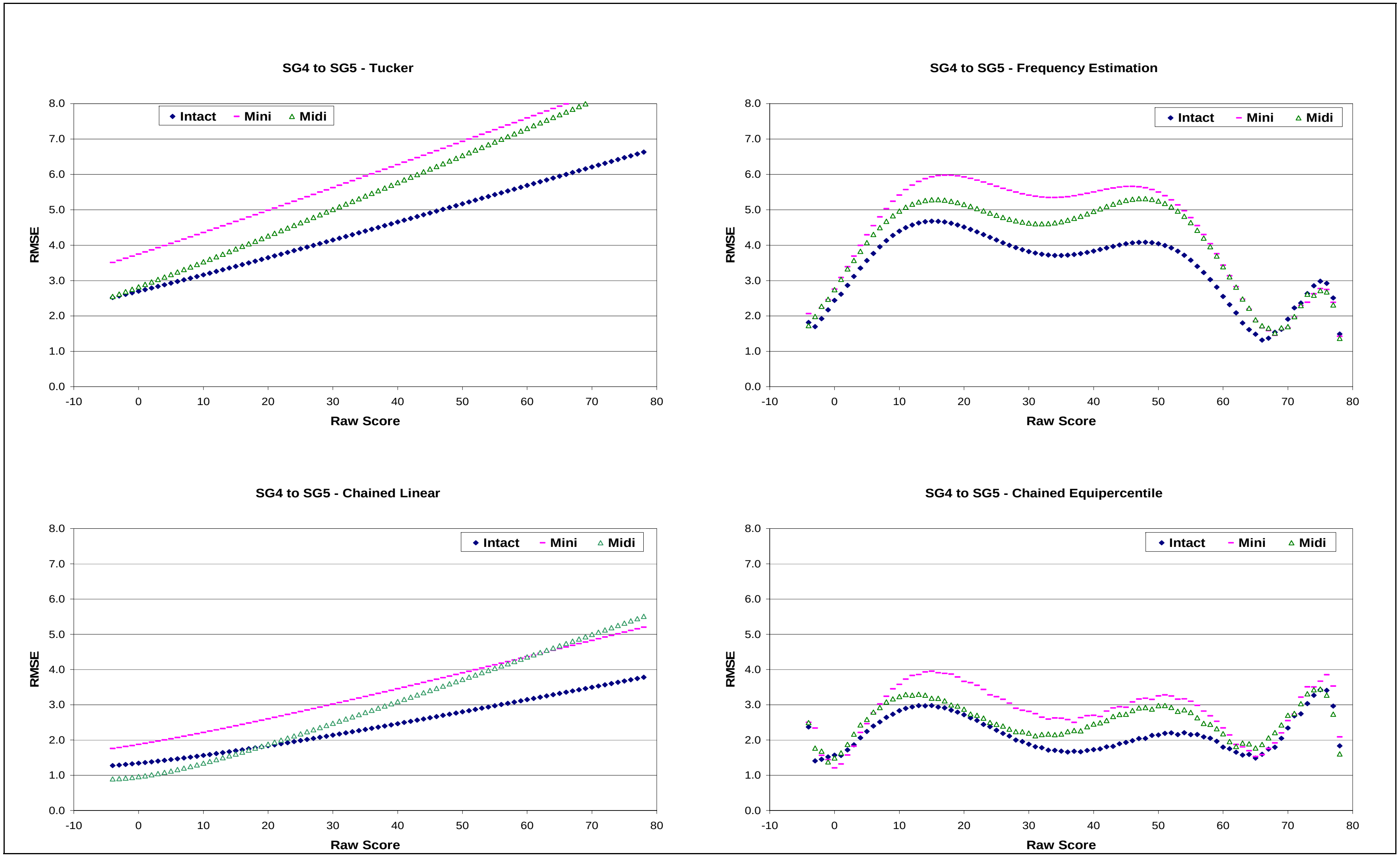


Figure 24. **Root mean squared error (RMSE) in the SG4-to-SG5 equating: *X2* to *Y2*.**

Table 10

Summary of Bias, Standard Errors of Equating (SEE), and Root Mean Squared Error (RMSE): X1 to Y1

	Tucker			Frequency estimation			Chained linear			Chained equipercentile		
	Intact	Mini	Midi	Intact	Mini	Midi	Intact	Mini	Midi	Intact	Mini	Midi
T-T												
Bias	0.42	0.49	0.53	0.43	0.49	0.54	0.40	0.48	0.52	0.41	0.48	0.55
SEE	0.14	0.16	0.16	0.19	0.21	0.21	0.14	0.17	0.16	0.21	0.23	0.23
RMSE	0.45	0.52	0.55	0.49	0.54	0.59	0.44	0.51	0.55	0.49	0.54	0.61
SG1 to SG2												
Bias	1.00	0.32	1.38	1.02	0.31	1.40	1.01	0.21	1.45	1.02	0.20	1.47
SEE	0.20	0.23	0.22	0.27	0.30	0.29	0.20	0.24	0.23	0.30	0.33	0.33
RMSE	1.03	0.39	1.40	1.06	0.44	1.44	1.03	0.34	1.47	1.08	0.40	1.52
SG3 to SG5												
Bias	1.14	2.10	1.04	1.24	2.12	1.32	0.51	1.45	0.56	0.55	1.51	0.74
SEE	0.35	0.40	0.39	0.48	0.52	0.52	0.38	0.44	0.42	0.55	0.60	0.62
RMSE	1.24	2.16	1.14	1.39	2.22	1.46	0.67	1.53	0.73	0.82	1.63	1.03
SG4 to SG5												
Bias	3.68	4.35	4.05	3.65	4.37	4.02	2.21	2.57	2.00	2.20	2.65	1.95
SEE	0.40	0.44	0.47	0.54	0.57	0.59	0.42	0.46	0.49	0.60	0.64	0.70
RMSE	3.70	4.38	4.08	3.69	4.40	4.06	2.25	2.61	2.06	2.28	2.72	2.08

Note. T-T = total group to total group.

Table 11

Summary of Bias, Standard Errors of Equating (SEE), and Root Mean Squared Error (RMSE): X2 to Y2

	Tucker			Frequency, estimation			Chained linear			Chained equipercentile		
	Intact	Mini	Midi	Intact	Mini	Midi	Intact	Mini	Midi	Intact	Mini	Midi
T-T												
Bias	0.50	0.78	0.39	0.45	0.77	0.34	0.53	0.87	0.41	0.49	0.86	0.36
SEE	0.18	0.20	0.20	0.24	0.27	0.26	0.19	0.22	0.21	0.27	0.30	0.30
RMSE	0.54	0.80	0.44	0.53	0.85	0.45	0.57	0.90	0.48	0.57	0.94	0.48
SG1 to SG2												
Bias	0.61	0.54	0.80	0.53	0.52	0.71	0.49	0.38	0.67	0.43	0.47	0.61
SEE	0.26	0.29	0.29	0.34	0.38	0.38	0.27	0.31	0.30	0.38	0.42	0.42
RMSE	0.66	0.62	0.85	0.65	0.70	0.82	0.56	0.52	0.74	0.61	0.67	0.76
SG3 to SG5												
Bias	1.07	1.60	1.14	1.24	1.95	1.41	0.47	0.83	0.24	0.87	1.27	0.75
SEE	0.52	0.58	0.57	0.69	0.73	0.73	0.57	0.64	0.63	0.80	0.87	0.88
RMSE	1.21	1.73	1.34	1.51	2.15	1.72	0.78	1.06	0.67	1.24	1.60	1.27
SG4 to SG5												
Bias	3.78	5.15	4.46	3.97	5.25	4.69	1.89	2.69	2.02	2.21	3.01	2.56
SEE	0.45	0.52	0.51	0.56	0.62	0.61	0.46	0.54	0.53	0.62	0.69	0.70
RMSE	3.81	5.18	4.50	4.03	5.30	4.75	1.95	2.75	2.11	2.33	3.14	2.69

Note. T-T = total group to total group.

SG3-to-SG5 equating. In the SG3-to-SG5 equating for both data sets, the mini anchor had the largest RMSE across most of the score range. The midi anchor and the intact anchor were intertwined (Figures 21 and 22). Either the midi anchor or the intact anchor had the smallest bias and RMSE on average, varying by the equating method, while the mini anchor yielded the largest bias and RMSE.

SG4-to-SG5 equating. Finally, in the SG4-to-SG5 group equating for data set 1 (Table 10 and Figure 23) and data set 2 (Table 11 and Figure 24), the intact anchor had the smallest equating bias and RMSE. Between the mini anchor and the midi anchor, the midi anchor produced smaller bias and RMSE.

Summary

The following list summarizes our findings from the operational data.

1 Anchor type effects on equating bias

1.1 Anchor types impact equating bias. Between the midi anchor and the mini anchor, the midi anchor performed better than the mini anchor across both the CE and PSE in both the SG3-to-SG5 equating and the SG4-to-SG5 equating. On the other hand, in the SG1-to-SG2 equating, the mini anchor outperformed the midi anchor. In the total-group-to-total-group equating, the mini anchor and the midi anchor each prevailed once.

1.2 In some cases, the midi anchor or the mini anchor even produce smaller equating bias than the intact anchor test, even though the latter has 75% more items.

1.3 Group differences substantially affect bias. The larger the ability differences between the two equating samples, the larger the bias. This finding is consistent with previous research (Dorans, Liu, & Hammond, 2008; Kolen, 1990; Sinharay & Holland, 2007).

1.4 Equating methods also affect the magnitude of the bias, conditioned on the group differences. When the two groups are very similar (e.g., the total group to total group), PSE tends to produce very similar or slightly larger bias than CE. When the group differences get large, the CE is less biased than the PSE. This finding is consistent with previous research (Sinharay & Holland, 2007; Wang, Lee, Brennan, & Kolen, 2008).

2 Anchor type effects on standard errors of equating

2.1 The anchor type impacts the SEE, although the differences across anchors are small. The intact anchor had the smallest SEE among the three anchor types; the midi anchor test had slightly smaller SEE than the mini anchor test.

2.2 Equating method impacts SEE. The PSE had slightly smaller SEE than CE; the curvilinear method had much larger SEE than its linear analogue.

3 Anchor type effects on total equating error (RMSE)

3.1 In general, the anchor type effects on RMSE are similar as those on equating bias: Whichever anchor reflects the standardized mean difference of groups of test-takers on the total tests most accurately will produce the most accurate RMSE.

3.2 The ability differences between the two groups substantially affect RMSE. The larger the ability differences, the larger the RMSE.

3.3 The CE and PSE interact with ability difference in their effect on RMSE. When the ability difference was near zero, PSE produced smaller RMSE; when the ability difference was not negligible, CE produced smaller RMSE.

Conclusions and Discussion

This study explores the use of a different type of anchor, a midi anchor that has a smaller spread of item difficulties than the tests to be equated, in contrast to the use of a mini anchor that is traditionally used in equating. Using SAT operational data, we built a midi anchor and a mini anchor. The impact of different anchors on observed score equating were evaluated and compared with respect to systematic error (bias), random equating error (SEE), and total equating error (RMSE). In order to examine how different types of anchors interact with differential group differences, we conducted equatings under four different scenarios where the differences between the old and new form samples varied between near zero to a large value.

The results show that the midi anchor generally produces more accurate equating results than the mini anchor with the following exceptions: (a) at the ends of the scale range in the curvilinear equatings and (b) in the SG1-to-SG2 equating. The SEE slightly favor the midi anchor. The overall bias, SEE, and the total equating errors when the midi anchor is used are either smaller than or very similar to those when the mini anchor test is used, except in the SG1-

to-SG2 equating. Our results are slightly different from those in Sinharay and Holland (2007), where they found that midi anchor tests consistently perform as well as the mini anchor tests.

Our findings suggest that a midi anchor test is preferred to a mini anchor test if equating accuracy at the top and at the bottom is not a primary concern. For example, if a licensure test has a cut score point somewhere around the middle of the score range, then using a midi anchor test will result in more accurate equating at that cut score point. On the other hand, if a test is a qualifying test in which the greatest accuracy is desired at the high end of the score range, then the use of a midi anchor test might not be optimal.

A very interesting phenomenon observed in this study is that the longer intact anchor that has the highest correlation with the total test does not always provide the least biased equating results. The shorter midi and mini anchors perform better than the intact anchor in certain situations. This finding is obviously not consistent with a common belief that a longer anchor with a higher correlation to the total test is supposed to function better than a shorter anchor that correlates at a lower level with the test to be equated.

It is not clear, though, which anchor performs most accurately under what circumstances. This lack of clarity might be related to the sampling variability or to the constructs being measured. For example, the mini anchor consistently outperformed the midi anchor in the SG1-to-SG2 equating, whereas the midi anchor consistently outperformed the mini anchor in the SG3-to-SG5 equating and in the SG4-to-SG5 equating. Or it might have something to do with group abilities on the verbal construct more than simple ability differences.

Certain limitations are found in this study. First, the study is based on only two data sets. Second, the mini anchor test and the midi anchor test are built based on an intact anchor. They might behave differently if they are administered as intact sections of their own. Third, the new form and the old form used in our study are essentially identical (one is the original order form and the other the scrambled form of SAT Verbal),[1] and the test-takers taking the new form and the old form represent different samples defined by demographic variables. Finally, only CE and PSE are investigated.

Further investigation should analyze more data sets from different testing programs, especially from a data set from a licensure testing program. Likewise, administering a midi anchor test operationally should be investigated. In addition, looking at equating accuracy with different anchor tests using IRT methods is worthy of further investigation.

References

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

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating* (ETS Statistical Rep. No. SR-98-02). Princeton, NJ: ETS.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81–97.

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

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2006). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 169–203). Amsterdam, Netherlands: Elsevier.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*, 97–104.

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

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education.

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an anchor test* (ETS Research Rep. No. RR-06-04). Princeton, NJ: ETS.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249–275.

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

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement, 32*(8), 632–651.

Notes

[1] Sinharay and Holland (2007) found in their extensive simulations that the difference in difficulty of the tests to be equated does not affect the equating performances of the anchors—so we do not think that this is a severe limitation of our study.