Listening. Learning. Leading.®

Research Report
ETS RR–11-09

Can Smoothing Help When Equating With Unrepresentative Small Samples?

Gautam Puhan

March 2011

Can Smoothing Help When Equating With Unrepresentative Small Samples?

Gautam Puhan

ETS, Princeton, New Jersey

March 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Dan Eignor

Technical Reviewers: Sooyeon Kim and Rick Morgan

Abstract

The study evaluated the effectiveness of log-linear presmoothing (Holland & Thayer, 1987) on the accuracy of small sample chained equipercentile equatings under two conditions (i.e., using small samples that differed randomly in ability from the target population *versus* using small samples that were distinctly different from the target population). Results showed that equating with small samples (e.g., $N < 50$) using either raw or smoothed score distributions can result in a substantial amount of random equating error (although smoothing reduced random equating error). Even with samples sizes of 100, the random equating error was quite large (greater than the difference that matters or DTM) for almost all score points. Moreover, when the small samples were unrepresentative of the target population, which is quite likely for small samples, the amount of equating bias (in addition to random equating error) was considerably large for both the raw and smoothed equatings. It was concluded that although presmoothing helped reduce random equating error, it is unlikely to reduce equating bias caused by using an unrepresentative sample. Other alternatives to the small sample equating problem that focus more on improving data collection than on improving existing equating methods are discussed.

Key words: log-linear presmoothing, test equating, small sample, random equating error, systematic equating error

Test equating involves two general sources of error (Kolen & Brennan, 2004). *Systematic equating error* (SEE) may result from factors such as using common items that do not adequately represent the total test in a nonequivalent anchor test (NEAT) design or using a sample that is clearly unrepresentative of the intended population to conduct the equating. *Random equating error* is present whenever a sample from a population of examinees is used to estimate the equating relationship. When the sample is large and representative of the population, then random equating error is small and the equating relationship is likely to be a close estimate of the equating relationship in the population. But when the sample is small and does not adequately represent the population, the equating can be inaccurate (Livingston, 1993). Testing programs frequently encounter small sample sizes and the number can often be as small as 20 or 30 test takers. Consequently, sample size can pose a problem when a new test form is introduced and has to be equated to an old form already on scale.

Several studies have examined the effect of small sample sizes on equating accuracy. For example, Kolen (1985) examined samples of 100 and 250 and found the standard error of equating to be sufficiently accurate with sample sizes of 250. Parshall, Du Bose Houghton, and Kromrey (1995) examined equatings based on 15, 20, 50, and 100 examinees and concluded that although equating bias was trivial even for samples as small as 15, random equating error substantially increased with smaller sample sizes. Skaggs (2005) evaluated equatings derived using sample sizes ranging from 15 to 200 in an equivalent groups design and concluded that for samples as small as 25, no equating was likely to be preferable, but for samples of 50 to 75, equating was preferable to no equating. In a recent study, Puhan, Moses, Grant, and McHale (2008) examined small sample linear equatings under the NEAT design. They found that for random equating error to be lower than the *difference that matters* or *DTM* (Dorans & Feigenbaum, 1994) criterion for scores within 1.5 standard deviations above or below the mean, at least 600 examinees for each of the new and old forms were needed.

Other studies have focused on modifying existing equating method to improve their performance with small samples. For example, Kim, von Davier, and Haberman (2006) proposed the synthetic linking function for small samples, which is defined as a weighted average of an estimated equating function (based on a small sample), and the identity function or no equating. They found that for samples as small as 10 or 25, the synthetic function was preferable to either no equating or using just the equating based on the small sample. Livingston and Kim (2009)

proposed the circle arc equating method for equating with small samples. This method constrains the equating curve to pass through two specified end points and an empirically determined middle point. The upper end point of the curve is determined by the maximum possible score on the test. The lower end point is fixed at the chance score (e.g., for a test consisting of 100, 4-choice multiple choice or MC items, the chance score is 25). The middle point is the equated mean score on the new form, which is determined by conducting a mean equating using the available small sample. These three points are used to determine a circle arc, which in turn, determines the equated scores on the new form. The authors found that the circle arc method outperformed other equating methods (e.g., Tucker, Levine, chained equipercentile, and chained linear) when the sample size was as small as 25 examinees.

Smoothing the data before conducting the equating (presmoothing) or after the equating function has been derived (postsmoothing) has also been shown to reduce error in small sample equating. For example, Livingston (1993), working with samples of 25, 50, 100, and 200, found that presmoothing significantly reduced equating error, particularly for the smallest size samples. He noted that presmoothing improved equating accuracy about as much as doubling the sample size would have done. Similarly, Hanson, Zeng, and Colton (1994) compared linear and identity equating (no equating) with unsmoothed, presmoothed, and postsmoothed equipercentile equating for five ACT assessments and found that smoothing improved equipercentile equatings when there were small samples, although there was no clearly preferred smoothing method when considering the pre- or postsmoothed methods. In a recent study (although not conducted to evaluate smoothing in small sample contexts), Cui and Kolen (2009) found that smoothing improved the estimation of the equating relationship by reducing total error (i.e., equating error and bias).

Although a proposed modification of an equating method may outperform existing equating methods when equating with small samples, it is possible that the proposed (modified) method produces equatings (in small sample situations) that are still not close enough to the true equating relationship. For example, consider a situation where the true equating relationship is known (i.e., an equating function derived using a very large data set). If we compare the small sample equating results based on an existing equating method (e.g., equating based on raw data or ER) versus a modified equating method (e.g., equating based on smoothed data or ES) and found that ES produced results closer to the true equating than ER, then ES would be considered

an improvement over ER. However, an important question is whether ES produced an equating that is close enough to the true equating to be considered appropriate for high stakes uses, such as admission or certification. Suppose, using the true equating function, a score of 80 on the new form converted to a score of 75 on the old form and using ER, the same score converted to 79.5 and using ES, the same score converted to a score of 79. Although using ES may be preferred over ER since it produced results closer to the truth, some may argue that the results of ES are still quite far from the true equating function and therefore not appropriate for high stakes use.

Purpose of the Study

This study evaluated the effectiveness of log-linear presmoothing (Holland & Thayer, 1987) on the accuracy of equatings with small sample sizes. Log-linear models provide useful smoothing techniques that allow the user to specify the number of moments of the observed distribution to be preserved in the smoothed distribution. This ensures that certain important properties of the observed data are retained in the smoothed data. If the observed sample is small, then the model may preserve only the mean and standard deviation of the observed distribution. If the observed sample is large, more moments such as the skewness and kurtosis of the observed distribution, can be preserved. Since the score distributions are smoothed before they are used for equating, the procedure is usually referred to as presmoothing.

Although the small sample size issue has been examined previously (e.g., Hanson, Zeng, & Colton, 1994; Livingston, 1993), the current study is distinct from previous studies in at least one important way. While previous studies used small samples that differed only randomly in ability from the target population, the current study will, in addition to evaluating equatings using small samples that differed randomly in ability, also evaluate equatings using small samples that are distinctly different from the target population. Although smoothing may reduce random error (Kolen & Brennan, 2004) in equating, the effectiveness of smoothing in reducing equating bias that may result from using unrepresentative small samples has not been investigated previously and was therefore evaluated in this study.

This added evaluation was motivated by advice from Holland, Dorans, and Petersen (2007), who stated that although smoothing may help in equating for moderate sample sizes, it may not be of much help for small samples, especially when it is unclear how well the small sample represents the intended population. Parshall et al. (1995) also pointed out that in actual testing programs the examinees taking a new test form may not represent a random sample but a

sample of convenience that may differ systematically from the population, and although the issue of sampling bias is not limited to small sample equatings, the statistical effects resulting from the use of a convenience sample are likely to be larger when the convenience sample is small.

Method

Data

Approximately 20,000 examinee responses from an operational form of a test, referred to as Test X, were used in this study. Form X consisted of 120 multiple-choice (MC) items measuring basic skills in elementary education and covered four subcontent areas or subscores (i.e., reading, mathematics, social studies, and science). Each subcontent area contributed equally to the total test (i.e., 30 out of 120 items or 25% of the total test).

The Approach

A hypothetical testing situation was created, whereby the single Form X was divided into two pseudo forms (i.e., Forms X1 and X2) with 84 items in each form and with 48 items common to both forms (see Figure 1 for illustration). As seen in Figure 1, Form X (consisting of 120 items) is divided into two alternate subforms, Forms X1 and X2, each consisting of 84 items. The shaded portion indicates the common section of 48 items between Forms X1 and X2. The items for Forms X1 and X2 were selected such that Forms X1 and X2 were also composed of the four content areas of reading, mathematics, social studies, and science, with each content area contributing to 25% of the total test. Throughout the study, Form X1 will be considered the new form and Form X2 will be considered the old form.

Equating Criterion

The hypothetical test design described above facilitated the computing of a criterion equating function. Since Forms X1 and X2 were created from one Form X (taken by 20,000 examinees), it essentially means that all the examinees took Form X1 and also Form X2. Therefore, Form X1 can be equated directly with Form X2 using a single group (SG) equating design. By using the SG design with such a large data set, one can be fairly confident that the resulting conversion is a very good approximation of the equating in the population. Livingston (1993) first proposed this approach for deriving a criterion equating function for the total test and it has been used in other equating studies (e.g., Puhan et al., 2008). Although the actual overlap

between Forms X1 and X2 is 48 items, only 24 of these common items were used as common items to conduct the small sample equatings, in order to mimic actual testing situations.

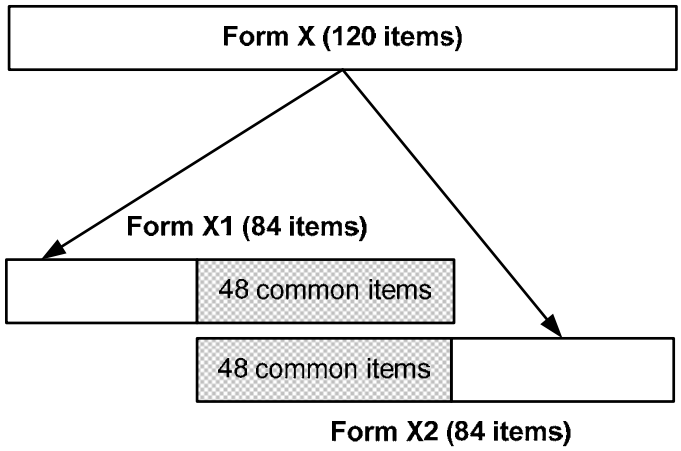

Figure 1. **Graph showing two alternate sub-forms (X1 and X2) created from one original Form X.**

The impact of presmoothing on equating was examined under two conditions, referred to as *Conditions 1* and *2*. Since the new and old forms were created from one form, 20,000 examinee responses were available for the new form and 20,000 examinee responses were available for the old form. In Condition 1, the assignment of examinee responses were unaltered (i.e., the new and old forms, each comprised 20,000 examinee responses). In Condition 2, an attempt was made to assign examinee responses to the new form that were not representative of the target population (i.e., the total data of 20,000 examinees). Using ethnicity as a background variable, 2,400 examinee responses from a less able ethnic group were assigned to the new form.[1] Similar to Condition 1, all 20,000 examinee responses were assigned to the old form. For testing programs dealing with small sample sizes, where the new form sample may be small and unrepresentative, it is often possible to accumulate data from different test administrations for the old form. Therefore it seemed reasonable to assign all 20,000 examinee responses to the old form in this condition. Although, for actual testing programs, it is unlikely that all examinees taking the new form belonged to one ethnic group, such an assignment of examinees was deliberately made in the current study to create a new form score distribution that was highly unrepresentative of the target population.

In both Conditions 1 and 2, small sample equatings were conducted with sample sizes of 25, 50, 100, and 500 examinees. Although 500 examinees may not necessarily be considered a small sample, it was included based on the previous finding by Puhan et al. (2008), who found that for random equating error to be lower than the DTM (Dorans & Feigenbaum, 1994) criterion for most scores within 1.5 standard deviations above or below the mean, at least 500 examinees each for the new and old forms were needed. It was also included to replicate (if possible) the finding by Livingston (1993) that smoothing offers the maximum benefit (i.e., lower equating error) for small samples and that the benefits diminish as the sample gets larger.

Evaluating the Effect of Presmoothing on Small Sample Equating

After the two testing conditions and sample sizes were determined, the study was conducted as follows. Although the three steps described below are specific to Condition 1, the same procedure was followed for Condition 2.

Step 1. To estimate the criterion equating function, Form X1 was equated to Form X2 (i.e., direct equipercentile equating without presmoothing of the two forms) using the SG equating design with the total data (N = 20,000). The resulting equipercentile conversion was considered as the criterion to which the small sample equatings were compared.

Step 2. For a particular condition (e.g., Condition 1 with sample sizes of 25), sample sizes of 25 examinee responses for the new and old forms each were drawn without replacement from the Form X1 and Form X2 data sets (i.e., N = 20,000 each for the new and old forms). Note that in Condition 2, the small samples for the new form were drawn from a specific subsample of 2,400 examinee responses and the small samples for the old form were drawn from the 20,000 examinee responses. Then three equatings (equating using raw data or ER, equating using presmoothed data preserving 2 moments or ES2, and equating using presmoothed data preserving 3 moments or ES3) were conducted to equate Form X1 to Form X2, making use of 24 common items and chained equipercentile equating (see Kolen & Brennan, 2004 and Livingston, 2004 for details on this equating method). In the N = 500 sample size condition, more moments could be preserved when smoothing the raw data. Therefore, three equatings (i.e., ER, ES3, and equating using presmoothed data preserving 5 moments or ES5) were conducted for this sample size.

Step 3. Step 2 was repeated 500 times and based on the repeated samples, the conditional standard error of equating (CSEE), weighted average of the CSEE (Avg SEE), conditional bias^2 (CBias), and weighted average bias (Bias) were computed. When computing the Avg SEE and

Bias, the CSEEs and CBias values were appropriately weighted using the raw proportion of examinees at each score point in the new form data. The root mean squared deviation (RMSD) was also calculated at the total test score level for evaluating the raw and smoothed small sample equatings. The RMSD is a useful statistic because it provides an estimate based on combining information from random and systematic error. Details on these statistical indexes are provided in Appendix B.

Although the different equatings are compared relative to each other, the practical criterion of the *DTM* was also used to evaluate the different equatings based on the raw or smoothed score distributions. The DTM is a unit equal to one half of a reporting unit. Because the tests used in this study had scores that progressed in 1-point increments, the DTM was defined as 0.5. Using a DTM criterion seemed reasonable because if a difference existed between the variability and accuracy indexes from the raw and smoothed equatings but the actual values were smaller than the DTM, then the differences are probably ignorable, as they may not result in a practical difference in the examinees' reported scores. On the contrary, if the differences between the variability and accuracy indexes from the raw and smoothed equatings were small but the actual values were substantially larger than the DTM, then the equatings based on either the raw or smoothed score distributions may be considered to be problematic.

Results

The summary statistics for Test X (New Form X1 and Old Form X2) are presented in Table 1. In Condition 1, the same group of 20,000 examinees took both Forms X1 and X2 and therefore their ability as indicated by the anchor score means is the same ($\overline{X}$ = 14.995). However, the new form total score is higher than the old form total score, indicating that the new form is easier than the old form. As evident from the anchor score in Condition 2, the new form sample is much less able than the old form sample. As seen in Table 1, the anchor-to-total test correlations for the new and old forms were fairly high in both conditions (min = 0.818 and max = 0.878). As mentioned earlier, although 48 items were common between Forms X1 and X2, 24 of those 48 common items were used as an anchor to conduct the equating to mimic realistic testing conditions where the anchor test length is usually about 20-25% of the total test length (see Kolen & Brennan, 2004, for detailed guidelines on anchor test length).

Table 1

Summary Statistics for New Form (NF), Old Form (OF), and Anchor in the Full Sample

N	Condition 1				Condition 2			
	20,000		20,000		2,400		20,000	
Score distributions (# of items)	NF total (84)	NF anchor (24)	OF total (84)	OF anchor (24)	NF total (84)	NF anchor (24)	OF total (84)	OF anchor (24)
Mean	54.672	14.995	53.407	14.995	45.175	12.430	53.407	14.995
SD	10.589	3.581	10.651	3.581	8.721	3.147	10.651	3.581
Anchor/total correlation	0.878		0.865		0.818		0.865	

The overall accuracy and variability of the equatings based on raw and smoothed score distributions and different sample sizes from Conditions 1 and 2 were estimated using the average SEE, bias, and RMSD indexes, which are presented in Tables 2 and 3, respectively. The conditional standard error of equatings or CSEEs and conditional bias or CBias estimates for the equatings based on raw and smoothed scores distributions from Conditions 1 and 2 are presented in Figures 2–5. The CBias associated with using the identity equating (i.e., no equating) function is also shown for both conditions.

Results for Condition 1

Average SEE, bias, and RMSD results. As seen in Table 2, in Condition 1 (i.e., where the small samples were randomly drawn from the target population of 20,000 test takers), the average SEE for the different equatings (i.e., ER, ES2, ES3, and ES5) are largest for the $N = 25$ sample size condition and became progressively smaller as the sample size increased. Within each of the $N = 25$, 50, and 100 sample size conditions, the average SEE was largest for ER and smallest for ES2. For the $N = 500$ sample size condition, the average SEE was largest for ER and smallest for ES3. The same pattern was observed for RMSD. The bias values for ER in the $N = 25$ sample size condition was somewhat large (0.97). This was larger than the bias values obtained for the raw and smoothed equatings from the remaining sample size conditions where the bias values were lower than the DTM of 0.5.

Table 2

Average SEE, Bias, and RMSD for Condition 1 Equatings

N	25			50			100			500		
Score distribution	Raw	ES2	ES3	Raw	ES2	ES3	Raw	ES2	ES3	Raw	ES3	ES5
Average SEE	4.14	2.20	2.65	2.92	1.51	1.78	2.01	1.26	1.55	0.91	0.53	0.68
Bias	0.97	0.09	-0.03	0.38	-0.01	0.00	0.10	0.03	-0.06	-0.02	0.02	0.00
RMSD	4.66	2.21	2.67	2.94	1.54	1.81	2.03	1.29	1.58	0.96	0.61	0.74

Conditional standard errors and bias. Although the average statistics described above provide a useful summary of random and systematic equating error, the CSEEs and conditional bias values are often considered more informative because they indicate the amount of variability and accuracy at each score point. Since CSEEs and CBias tend to be less stable at score points with less data, it was decided to focus on the score points between the 5th and 95th percentiles because most of the data was observed within this score range. For the new form data (N = 20,000), the 5th and 95th percentiles were score points 38 and 72 and the CSEEs and CBias values were evaluated for scores within this range. For Condition 1, this information is provided in Figures 2 and 3. In Figures 2 and 3 and the remaining figures (i.e., Figures 4, 5, and 6), raw indicates ER, smoothed (2) indicates ES2, smoothed (3) indicates ES3, smoothed (5) indicates ES5, and the straight dashed line indicates the DTM of 0.5. As seen in Figure 2, the CSEEs for N = 25 sample size condition are the largest and become smaller as sample size increases. (i.e., N = 50, 100, and 500, respectively). Within each sample size condition, the CSEEs for ER are the largest. For the N = 25, 50, and 100 sample size conditions the CSEEs for ES2 are slightly smaller than those for ES3 and for the N = 500 sample size condition, the CSEEs for ES3 are slightly smaller than those for ES5 (especially around the middle of the score distribution). Also, for the N = 500 sample size condition, the CSEEs for ER, ES3 and ES5 were more similar to each other than what was observed in the smaller sample size conditions. Since the purpose of smoothing is to estimate the score distribution that would occur in a much larger group of examinees, it is not surprising that as the sample size increased, the benefits of smoothing in terms of reducing random equating error diminished. As seen in Figure 3, the conditional bias values for ER, ES2, ES3 (for the N = 25, 50, and 100 sample size conditions) and ER, ES3 and ES5 (for the N = 500 sample size condition) are close to zero around the middle of the distribution. Since the small samples differed randomly in

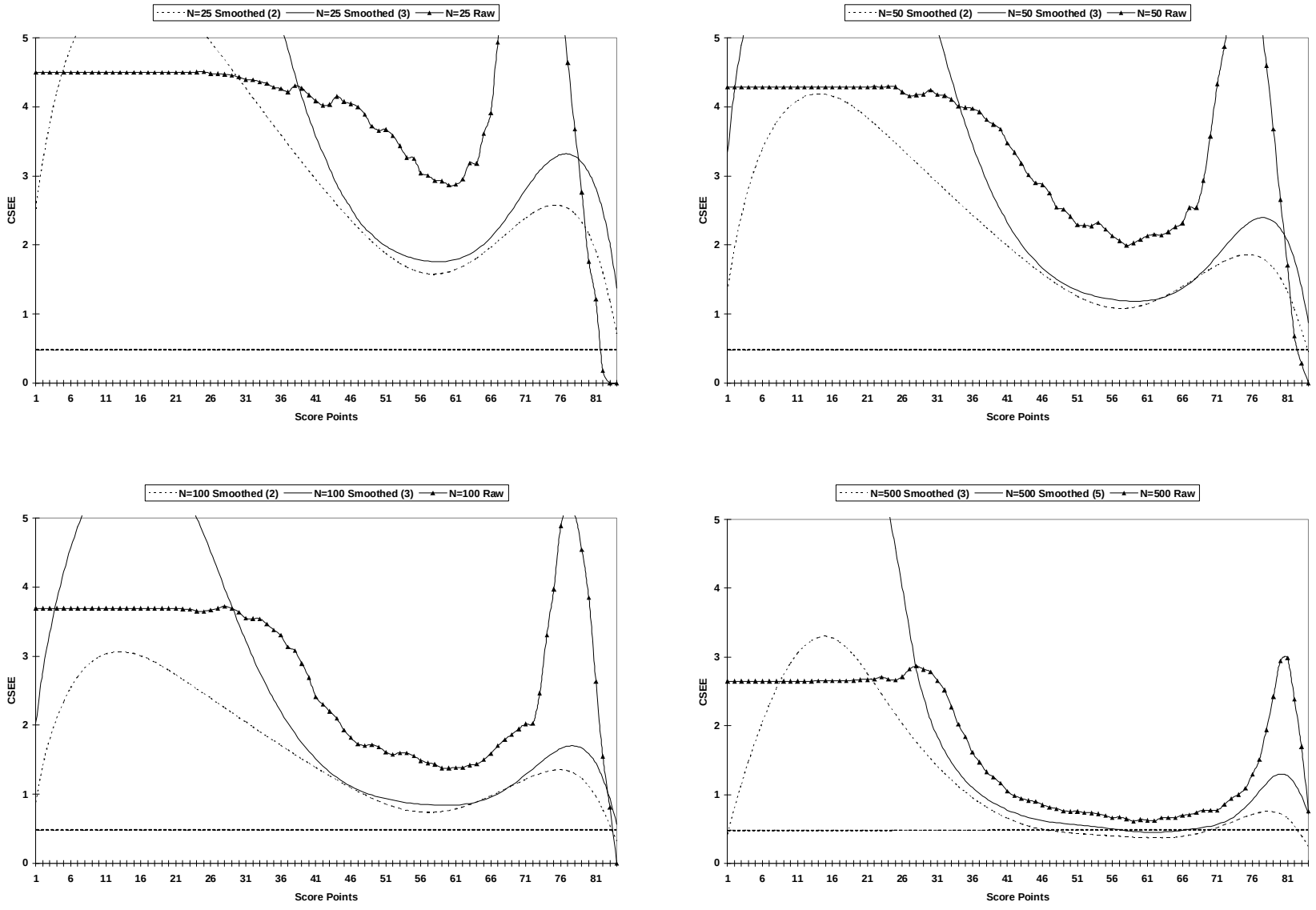

Figure 2. **Conditional standard error of equating (CSEE) based on raw and smoothed score distributions (Condition 1).**

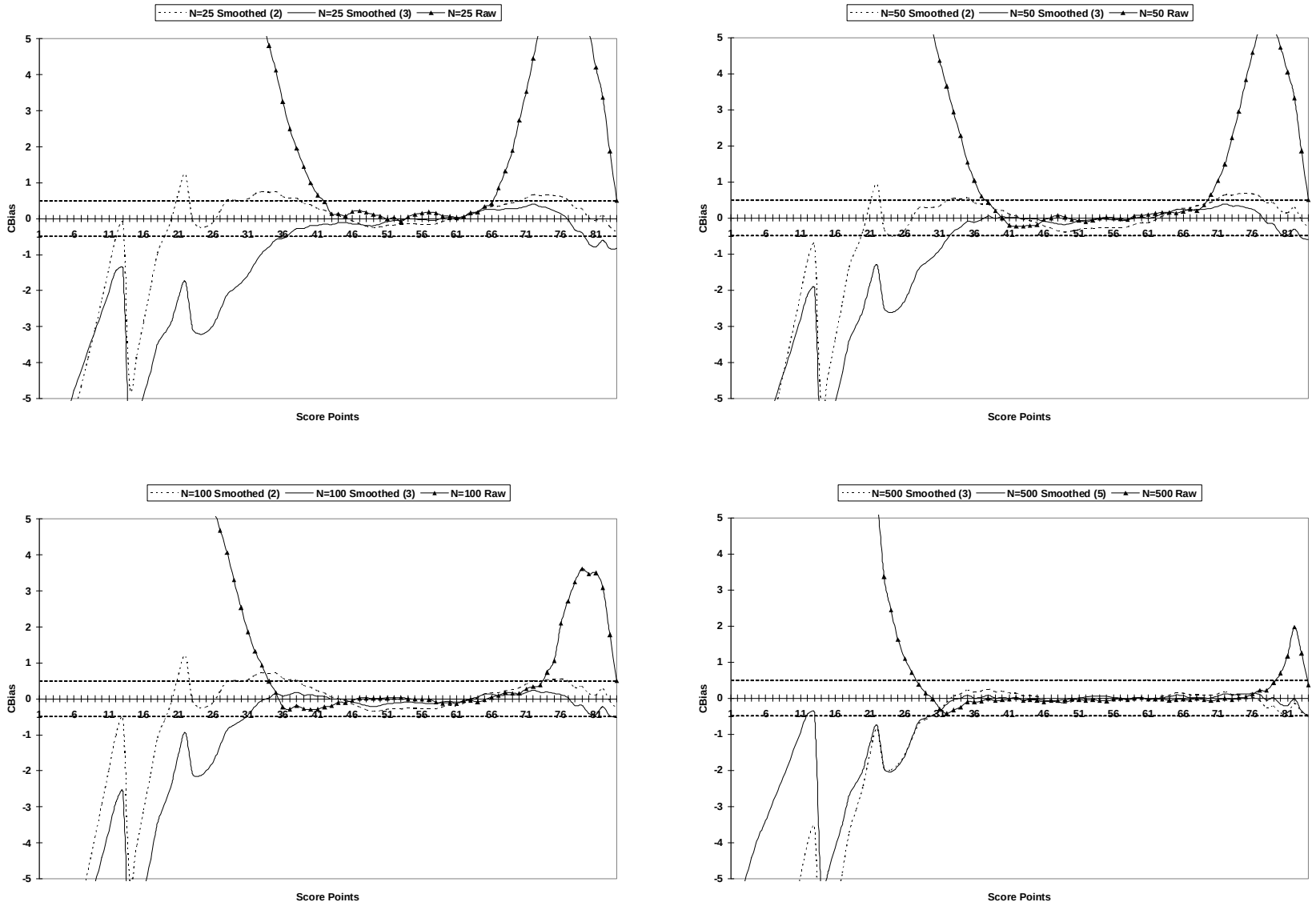

Figure 3. **Conditional bias (CBias) based on raw and smoothed score distributions (Condition 1).**

ability from the target population and positive and negative bias values tend to cancel each other, such an outcome was not completely unexpected (see Parshall, Du Bose Houghton, & Kromrey, 1995, who found similar CBias values in a small sample equating study where the small samples were random samples from the target population).

Results for Condition 2

Average SEE, bias, and RMSD results. As seen in Table 3, in Condition 2 (i.e., where the small samples were randomly drawn from the sub samples that were highly unrepresentative of the target population), the average SEE for the different equatings (i.e., ER, ES2, ES3, and ES5) are largest for the N = 25 sample size condition and became progressively smaller as the sample size increased. Within each of the N = 25, 50, and 100 sample size conditions, the average SEE was largest for ER and smallest for ES2. For the N = 500 sample size condition, the average SEE was largest for ER and smallest for ES3. The same pattern was observed for RMSD. Unlike Condition 1, where the bias values for most equatings was lower than the DTM of 0.5, in this condition, the bias values are much larger than the DTM. The bias values for ER, ES2, ES3, and ES5 for most sample sizes were about 2, indicating that if the sample used to conduct the equating is highly unrepresentative of the target population (which is quite likely in small sample size scenarios), there can be considerable amount of equating bias in addition to large standard error of equating. The bias value also indicates that if the sample used to conduct the equating is unrepresentative of the target population, then systematic bias will exist not only for the small sample sizes (e.g., N = 25) but also for the larger sample sizes (e.g., N = 500).

Table 3

Average SEE, Bias, and RMSD for Condition 2 Equatings

N	25			50			100			500		
Score distribution	Raw	ES2	ES 3	Raw	ES2	ES 3	Raw	ES2	ES 3	Raw	ES3	ES 5
Average SEE	4.57	2.40	3.05	3.29	1.71	2.07	2.26	1.16	1.46	1.01	0.61	0.77
Bias	3.21	1.93	2.02	2.46	1.99	1.90	2.20	2.04	1.98	1.92	1.98	1.95
RMSD	6.02	3.13	3.77	4.42	2.67	2.93	3.34	2.40	2.60	2.33	2.22	2.23

Conditional standard errors and bias. For Condition 2, the CSEEs and CBias values are provided in Figures 4 and 5. The 5th and 95th percentiles for the new form data (N = 2,400) were 33 and 62 and the CSEEs and CBias values were evaluated for scores within this range. As seen in Figure 4, the CSEEs in Condition 2 followed a similar pattern as was observed in Condition 1 (i.e., the CSEEs are largest for the N = 25 sample size and become smaller for the larger sample size conditions). Within each sample size condition, the CSEEs for ER are the largest. For the N = 25, 50, and 100 sample size conditions, the CSEEs for ES2 are slightly smaller than ES3 and for the N = 500 sample size condition, the CSEEs for ES3 are slightly smaller than ES5 (especially around the middle of the score distribution). Also for the N = 500 sample size condition, the CSEEs for ER is closer to the CSEEs for ES3 and ES5. As seen in Figure 5, the conditional bias values for ER, ES2, ES3 (for the N = 25, 50, and 100 sample size conditions) and ER, ES3, and ES5 (for the N = 500 sample size condition) are quite large for most score points between 33 and 62. Since the small samples were drawn from a subsample that was highly unrepresentative of the target population, high CBias values were expected.

Often in very small sample situations, testing programs choose not to equate on the basis of the very small sample. Instead, they report unequated scores by making a strong assumption that the old and new forms are the same in difficulty. After the new form has been administered several times and enough data is collected, equating is conducted using the accumulated sample and scores on the new form are then reported (without going back and rereporting) on the newly equated scale. The important question that arises in such situations is whether it is reasonable to assume that the new and old forms are the same in difficulty and therefore using the identity equating function (i.e., reporting unequated raw scores) for reporting scores on the new form. To answer this question for the test used in this study, the CBias values that would result by using the identity equating function was plotted alongside the CBias values from ER, ES2, and ES2 from the N = 25 sample size condition. As seen in Figure 6, the CBias from using the identity equating function would be much smaller than the CBias from using an unrepresentative sample to conduct the equating, indicating that in this case, it may be better not to equate than to equate with a highly unrepresentative equating sample.

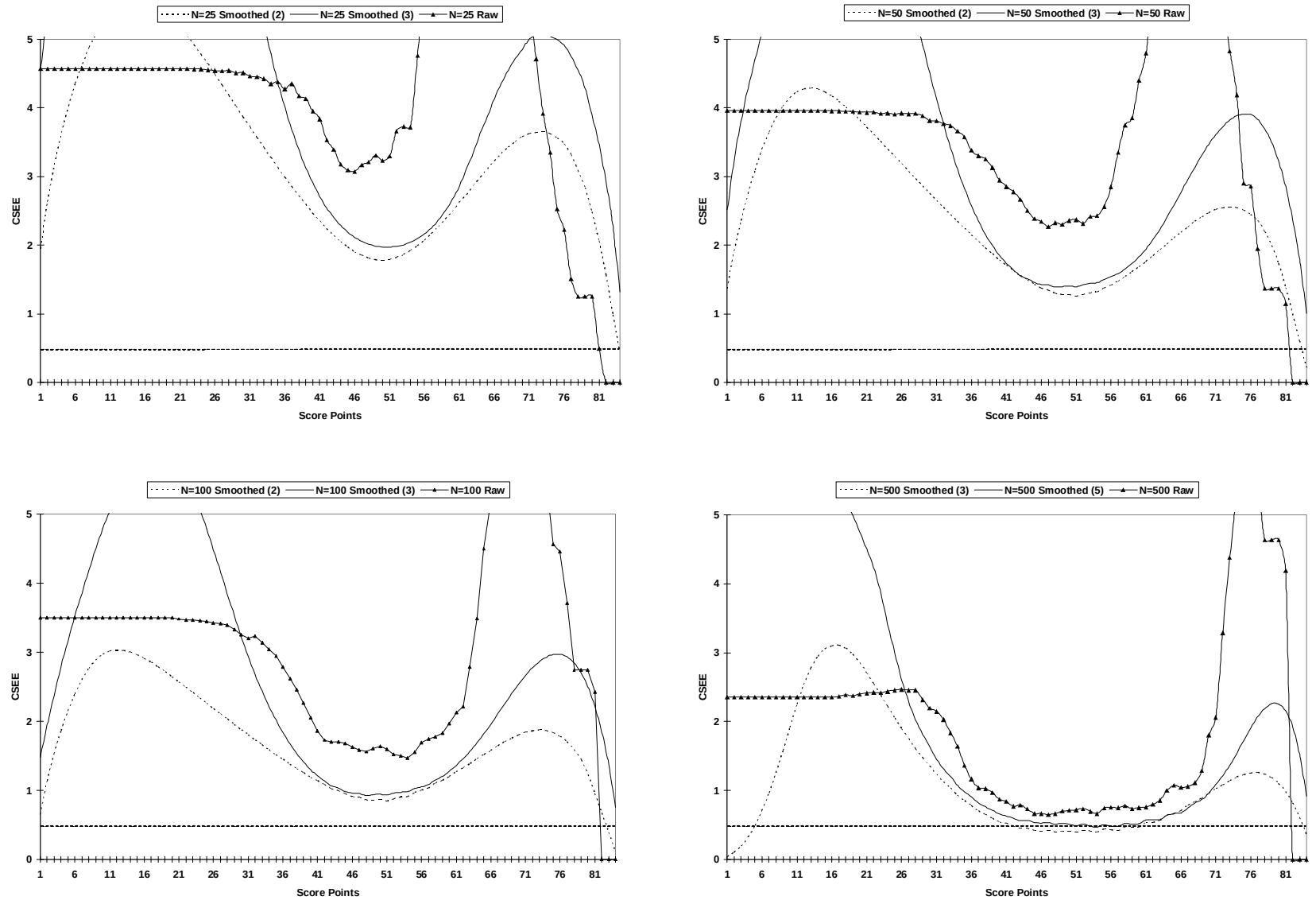

Figure 4. **Conditional standard error of equating (CSEE) based on raw and smoothed score distributions (Condition 2).**

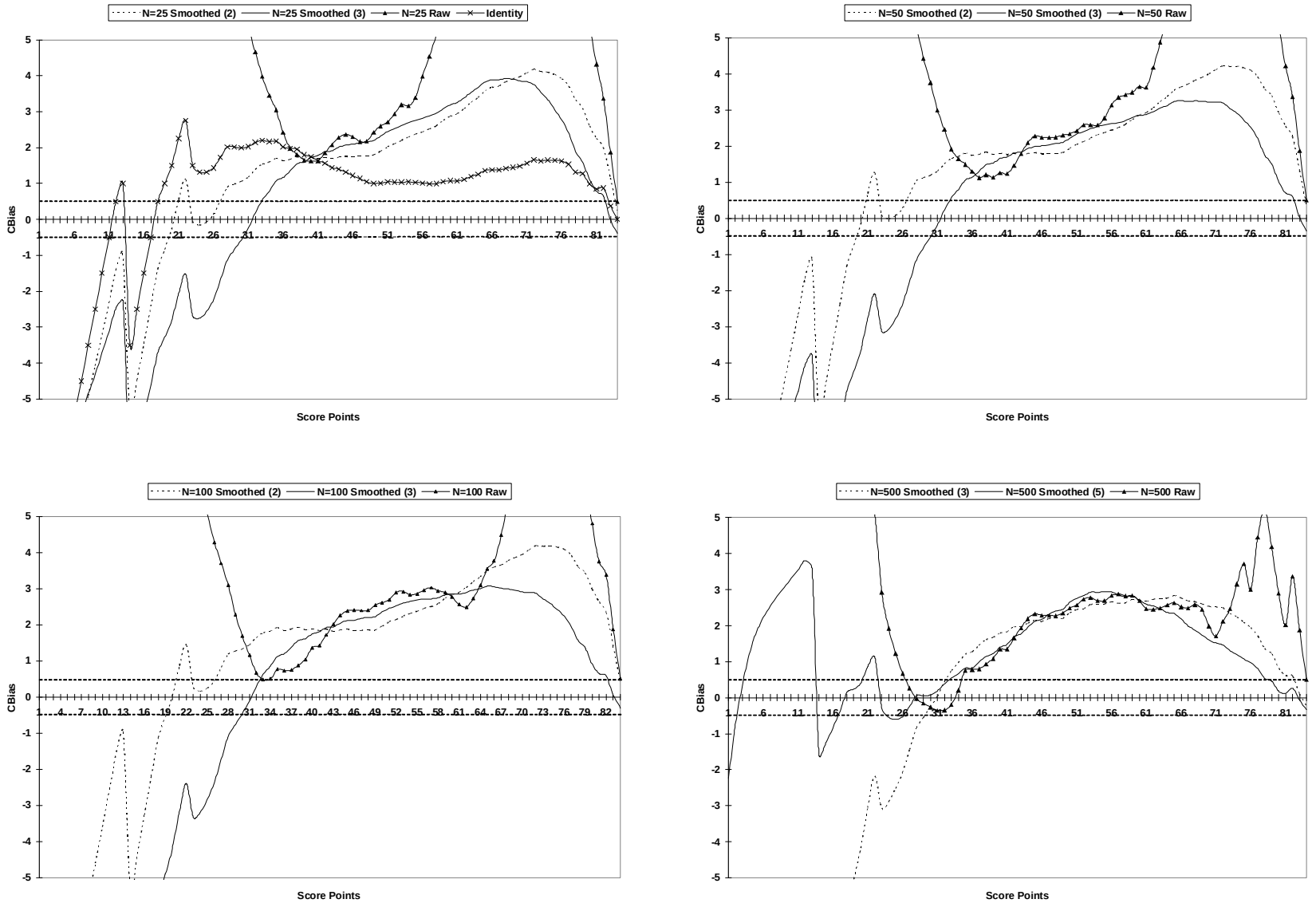

***Figure 5*. Conditional bias (CBias) based on raw and smoothed score distributions (Condition 2).**

Additional Conditional Standard Errors of Equating (CSEE) and Conditional Bias (CBias) Results

In Condition 2, the new form sample was selected to be highly unrepresentative of the target population. Some further analyses were conducted using new and old form samples that were still unrepresentative of the target population but maybe more similar to what is observed in actual testing conditions. The full data of 20,000 test takers were actually accumulated over four test administrations where the same form (i.e., Form X) was administered without any modification. Therefore, the mean score for each group of test takers was used to identify the most and least able groups from the four groups of test takers. Furthermore, using ethnicity as a background variable, some strong test takers were removed from the less able group and some weak test takers were removed from the more able group to further widen the ability gap between these two groups (standardized mean difference [SMD] between the two groups on the 24 anchor items was 0.36). This seemed reasonable because new and old form samples in actual testing programs can sometimes differ by much as 0.3 SMD or higher. This is especially likely when a new state adopts an existing test title for licensure purposes and the new test-taking sample is very different in ability than the existing test-taking sample. The less able group (N = 4,000) was assigned to the new form and the more able group (N = 6,000) was assigned to the old form. The resampling study was then conducted using the same steps that were followed for conditions 1 and 2. The CSEE and CBias values obtained from this condition for one sample size (i.e., N = 25) are provided in Figure 6. As seen in Figure 6, the CSEEs are still considerably large (greater than the DTM) around the middle of the score distribution. The CBias is also much larger than 0.5 for scores in the middle of the distribution. Although the CBias values are slightly smaller than what were observed in Condition 2, they are still quite large, indicating an inaccurate equating. As was observed in Condition 2, the CBias resulting from the use of the identity equating function is smaller than the CBias resulting from using the small sample to conduct the equating.

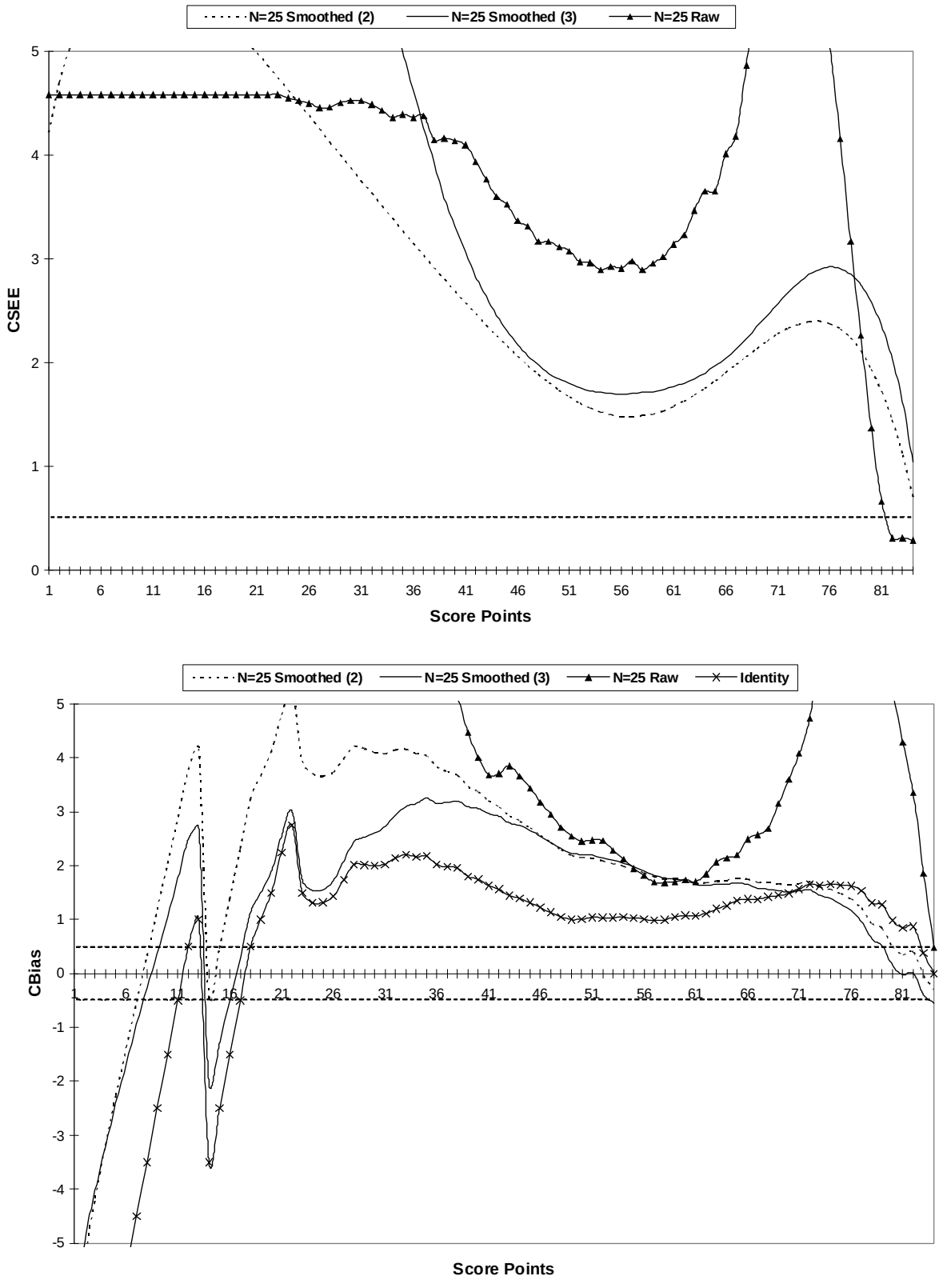

***Figure 6*. Conditional standard error of equating (CSEE) and conditional bias (CBias) based on raw and smoothed score distributions (additional analyses).**

Discussion

This study evaluated the effectiveness of log-linear presmoothing in improving the accuracy of equatings with small sample sizes. In Condition 1 the small samples were drawn randomly from the total available data (target population) and therefore only differed randomly in ability from the target population. In Condition 2 the small samples were drawn in a manner that caused them to differ systematically from the target population. According to Kim and Dorans (2008), the chances of obtaining a representative sample from the population is greatly diminished if only a small sample was available. Therefore, Condition 2 was included to evaluate the effectiveness of smoothing in reducing equating bias that may result from using unrepresentative small samples to equate.

In Condition 1, the average and conditional random equating error values were largest in the smallest sample size condition (N = 25) and became progressively smaller with larger sample sizes. Within each sample size condition, the equatings based on the raw score distributions had the largest average and conditional random equating error and the equatings based on smoothed score distributions that preserved fewer moments tended to have less random equating error. Similar to earlier studies, the benefit of smoothing (in terms of reduced standard error of equating) was largest for the smaller samples as compared to the larger samples. The average bias and conditional bias values (in the middle of the score distribution) were smaller than the DTM for the raw and smoothed equatings for all sample sizes, except the raw equating in the N = 25 sample size condition where the bias was larger than the DTM.

In Condition 2, the average and conditional random equating error were [in the smallest sample size condition (N = 25) and became progressively smaller with larger sample sizes. The equatings based on the raw score distributions had the largest average and conditional equating error and the equatings based on smoothed score distributions that preserved fewer moments tended to have less random equating error. However, unlike Condition 1 where the equating bias was quite small for most sample sizes, the equating bias in this condition was quite large. The overall bias was about 2 for most sample sizes for both the raw and smoothed equatings. The conditional bias was also quite large where most of the data was observed. In fact, the conditional bias resulting from using the identity equating (i.e., no equating) was smaller than the conditional bias resulting from equating with unrepresentative raw or smoothed small samples. Since all the samples (N = 25, 50, 100, and 500) in this condition were systematically different

from the target population, large average and conditional bias were observed for all sample size conditions. Additional analyses with unrepresentative samples also resulted in a considerable amount of equating bias. In this case also, the identity equating resulted in smaller CBias than equating with small unrepresentative raw or smoothed samples.

Conclusion and Recommendation

As evident from the results of this study, equating with very small samples irrespective of whether they are representative or unrepresentative of the target population can result in substantial amount of random equating error. Even with sample sizes of 100, the random equating error was quite large for the entire score region that was examined. This is in agreement with Kolen and Brennan (2004), which suggested that the use of the identity may be preferable to using an equating method, especially with sample sizes at or below 100 examinees per test form. Moreover, if the small sample used to conduct the equating is unrepresentative of the target population, which is more likely for small samples, it can result in substantial equating bias. Although smoothing may help in reducing random equating error, it is unlikely to reduce equating bias resulting from using unrepresentative samples and this, as Kim, von Davier, and Haberman (2006) pointed out, can counteract the gain due to reduction in the standard error of equating. If testing programs are already equating with small samples (based on raw data), then presmoothing the score distributions and using the smoothed distributions to conduct the equating may provide an improvement (especially in terms of reduced random error) over the current procedure. However, as evident from the results of this study, equating using small samples (based on smoothed data) can still produce a very inaccurate equating, which may not be adequate for high stake uses such as admission or certification. Therefore other approaches (examples are provided in a later section) to deal with the small sample equating problem are perhaps warranted.

What then is the answer to the small sample equating problem? As reviewed earlier in the literature section, methodological approaches to deal with this problem exist in the equating literature. Some examples include (a) small sample equating with log-linear smoothing (Livingston,1993); (b) synthetic linking function (Kim, von Davier, & Haberman, 2006), which is a weighted average of an estimated equating function (based on a small sample) and the identity function or no equating; and (c) the circle arc method (Livingston & Kim, 2008), which constrains the equating curve to pass through two specified end points and an empirically

determined middle point. However, such approaches cannot be expected to produce accurate equating results if the data is unrepresentative of the target population (i.e., may result from sampling bias). Although log-linear presmoothing has been shown to improve equating in small samples, it is still a matter of debate if it produces accurate results in very small samples. According to Holland, Dorans, and Petersen (2006) and Petersen (2007), smoothing helps in equating for moderate sample sizes but may not be of much help for very small samples. This is especially true when it is unclear how representative the small sample is of the intended population. As evident, the results of this study (i.e., large equating bias in Condition 2) support this assumption. Similarly, if one were to use the synthetic linking function for the data used in Condition 2, then using the synthetic linking function (which is a weighted average of an estimated equating function based on a small sample and the identity function) may help reduce equating bias but the bias could still be quite large. The conditional bias in Condition 2 (see Figure 5) for the synthetic linking function will probably be somewhere between the conditional bias for the identity and the conditional bias for the small sample equating, which is still less than ideal. Finally, if the circle arc method was used to equate using the data from Condition 2, then it would also most likely produce inaccurate equating results. As seen in Livingston and Kim (2008, p. 12), the equating bias for small samples, especially in the middle of the score range, are quite similar for the circle arc methods and the chained equipercentile method with presmoothing. Therefore it is reasonable to expect that if the circle arc method was used in this study, it would have resulted in a similar amount of equating bias (i.e., large) as was observed for the chained equipercentile equatings using smoothed data in Condition 2. The above examples along with the results of this study suggest that methodological approaches to equating may not always lead to an accurate small sample equating, especially when the small sample is highly unrepresentative of the population.

So how do we equate when the available sample is small and likely unrepresentative of the population? Since proper data collection is regarded as the most important aspect of any equating (Holland, Dorans, & Peterson, 2007), an equating design whereby data conducive to improved equatings can be collected may be the solution to the small sample equating problem. An example of such a design is the single group nearly equivalent test or SiGNET design (Grant, 2006; Puhan et al., 2008). The basis of this design is that examinees take two largely overlapping test forms within a single administration. The scored items for the operational form are divided

into mini tests called testlets. An additional testlet is created but not scored for the first form. If the scored testlets are Testlets 1-6 and the unscored testlet is Testlet 7, then the first form is comprised of Testlets 1-6 and the second form is comprised of Testlets 2-7 and Testlets 2-6 are common to both test forms. They are given as a single administered form and when a sufficient number of examinees have taken the administered form for a SG equating, the second form (Testlets 2-7) is equated to the first form (Testlets 1-6) using SG equating. As evident, there are at least two merits of the SiGNET design over the nonequivalent anchor test or NEAT design. First, it facilitates the use of a SG equating design which has the least random equating error of all designs, and second, it allows for the accumulation of data to equate the second form with a larger sample, which is more likely to be representative of the target population. Since the examinees scores are based on only the first form (i.e., the operational form), the two forms can administered until sufficient data is collected to equate the second form. Some may argue that having a large overlap between the new and old forms and delivering the same form in repeated administrations to accumulate enough data for equating under the SiGNET design increases the risk of exposure. In reality, however, since this design is proposed for very small volume tests with test takers often testing in different parts the country, the risk of overexposure may be minimal as compared to high volume tests where even though new forms are introduced more frequently, there is still some overlap of items (i.e., anchor items) which are exposed to much larger testing groups. Furthermore, as Kim and Dorans (2008) point out, test security may be less of an issue when there is little financial incentive to steal tests that can be sold to only a few people.

Another alternative may be to increase the number of common items in the new and old forms, which is similar to the SiGNET model. But unlike the SiGNET model where two forms are administered as a single form to allow for a SG equating design, this design still uses the regular NEAT model but uses a large number of common items which may help in reducing random equating error (see Puhan, 2010, which showed empirically that increasing the number of common items resulted in the lowering of random equating error).

Other alternatives which are more policy rather than measurement driven may be useful when addressing the small sample equating problem. One such alternative (if contractual obligations and state laws permit) would be to administer the small sample tests in fewer testing administrations throughout the year, thereby increasing the sample size per administration, which

may help with the small sample equating problem. A second alternative may be to report unequated scores and then state that these scores cannot be compared across test forms. However, after the new form has been administered several times and enough data is collected, equating can be conducted using the accumulated (larger) sample and scores on the new form can then be reported on the newly equated scale. Such an approach, according to Kim and Dorans (2008), would protect testing organizations from issues due to the use of an unstable or inaccurate equating function resulting from using small sample sizes to conduct the equating.

Limitations and Future Research

In this study, it was not clearly defined how much small samples should differ from the target population for them to be considered unrepresentative. Future studies may use data from different testing programs to gain a more realistic understanding of how much do small samples actually differ from a larger target sample. For example, if a small sample testing program used the same test form during several test administrations (although this maybe undesirable due to risk of overexposure of the form) then the data across the different test administrations can be accumulated and considered as the larger target sample. Then, data from each administration can be compared to this target sample (e.g., comparing frequency distributions, means, etc) to determine how much they actually differ from the target sample. Based on such information, future studies can systematically vary the degree of *unrepresentativeness* of small samples from the target population and then evaluate its impact on equating. Small samples, especially for licensure tests, can fluctuate considerably when new user states adopt or existing user states stop using a test title for licensure purposes. Since such fluctuations can contribute to equating bias, future studies should also evaluate if small sample equatings are invariant to such changes in the test-taking sample.

References

Cui, Z., & Kolen, M. (2009). Evaluation of two new smoothing methods in equating: the cubic B-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement, 46*(2), 135-158.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Hanson, M. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report No. 94-4). Iowa City, IA: American College Testing.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Program Statistics Research Technical Report No. 87-79). Princeton, NJ: ETS.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2006). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 169-201). Amsterdam, the Netherlands: Elsevier.

Kim, S., & Dorans, N. J. (2008). *Linking scores with small samples of examinees: A review of methodological and data collection approaches*. Unpublished manuscript.

Kim, S., von Davier, A. A., & Haberman, S. (2006). *Equating with small samples* (ETS Research Report No. RR-06-27). Princeton, NJ: ETS.

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9,* 209-223.

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

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30,* 23-29.

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

Livingston, S.A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46(*3), 330-343.

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

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54-75.

Puhan, G., Moses, T., Grant, M., & McHale, F. (2008). *An alternative data collection design for equating with very small samples* (ETS Research Report No. RR-08-11). Princeton, NJ: ETS.

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

Notes

[1]Figure A1 in the appendix shows how much the subsample differed from the full data. As evident from the graph, although smoothing helped in removing some of the irregularities of the subsample, it still retained its basic shape. In other words, smoothing is unlikely to overcome the bias that may result from using an unrepresentative sample to conduct the equating.

[2] In this study (especially in Condition 2), systematic error in the equating may be observed because of using samples that are unrepresentative (i.e., sampling bias) of the target population to conduct the equating. Such systematic error may not fit the usual definition of equating bias which is considered as systematic error arising from factors such as using an inadequate equating method or using an anchor test that not represent the total test in content and difficulty, etc. However, for the sake of convenience, we refer to this inaccuracy as equating bias throughout the paper.

Appendix A

Score Distribution of Full Sample

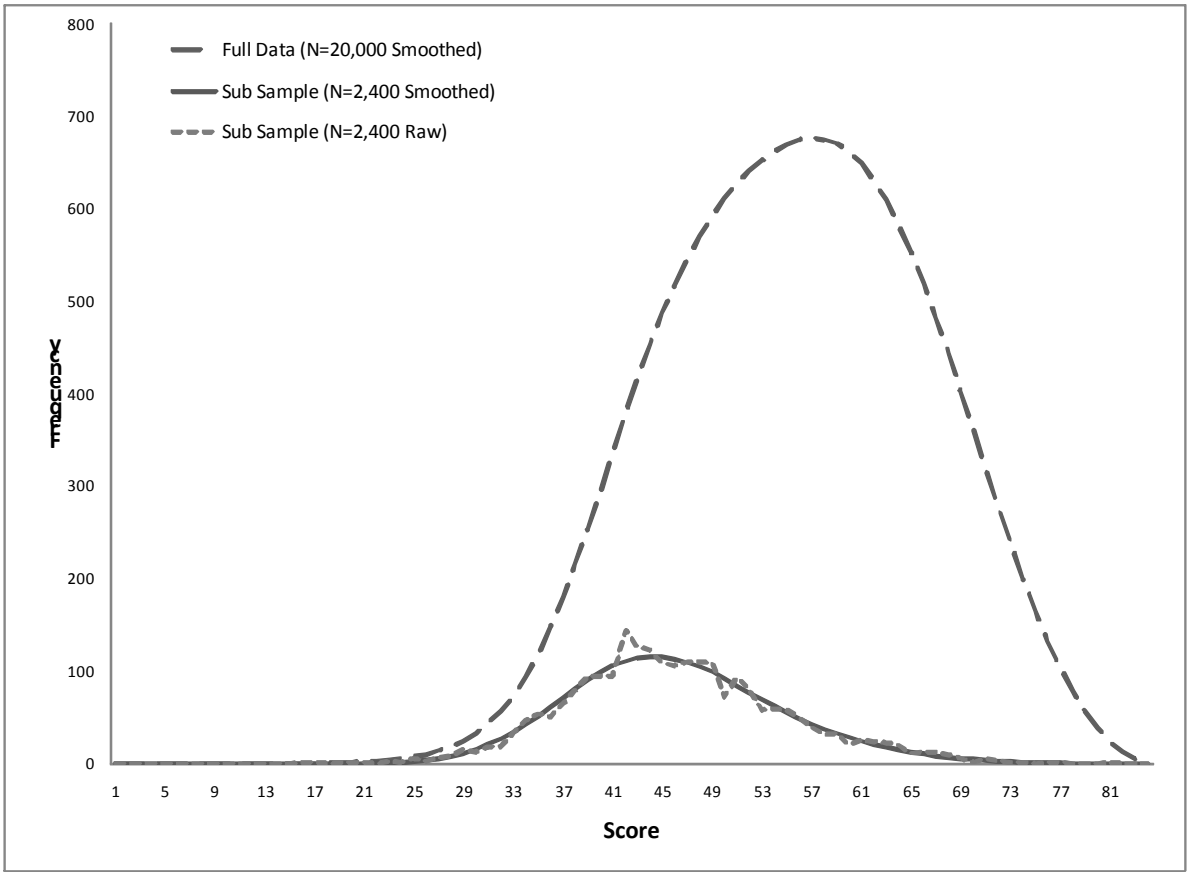

Figure A1. **Score distribution of full sample (smoothed) and subsample (raw and smoothed).**

Appendix B

Conditional Standard Errors of Equating

The formula for CSEE is

$$CSEE_j = \sqrt{\frac{1}{I}\sum_i \left(\hat{e}_y(x_{ij}) - \frac{\sum_i \hat{e}_y(x_{ij})}{I} \right)^2},$$

where I is the number of replications in the simulation (i = 1 to I, and I = 500) and $\hat{e}_y(x_{ij})$ is the equated score at score = x_j estimated for replication = i. The formula for Avg SEE is

$$Avg\ SEE = \sqrt{\sum_j p_j CSEE_j^2},$$

where p_j is the raw proportion of examinees at each score point in the new form data. The formula for CBias is

$$CBias_j = \frac{1}{I}\sum_i \left(\hat{e}_y(x_{ij}) - e_y(x_j) \right),$$

where I is the number of replications in the simulation (i = 1 to I, and I = 500), $e_y(x_j)$ is the criterion single group equated score at score = x_j and $\hat{e}_y(x_{ij})$ is the equated score at score = x_j estimated for replication = i. The formula for Bias is

$$Bias = \sum_j p_j CBias_j,$$

where p_j is the raw proportion of examinees at each score point in the new form data. The formula for RMSD is

$$RMSD = \sqrt{AvgBias^2 + AvgSEE^2},$$

where $AvgBias^2$ is the sum of the squared conditional bias values weighted by the raw proportion of examinees at each score point in the new form data. The formula is

$$AvgBias^2 = \sum_j p_j CBias_j^2.$$