*Center for Advanced Studies in Measurement and Assessment*

*CASMA Research Report*

*Number 54*

# Impact of Degrees of Postsmoothing on Long-Term Equated Scale Score Accuracy

*Stella Y. Kim*[†]
*YoungKoung Kim*[*]
*Tim Moses*[*]

December 2020

---

[†]Stella Kim is Assistant Professor in Educational Research, Measurement, and Evaluation at University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223-0001 (email: stella-kim@uncc.edu).

[*]YoungKoung Kim is Senior Psychometrician at the College Board (email: ykim@collegeboard.org). Tim Moses is Brennan Chair of Psychometric Research at the College Board (email: tmoses@collegeboard.org).

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

# Contents

**1 Introduction** **1**

**2 Cubic-Spline Postsmoothing** **2**

**3 Method** **4**
3.1 Data Generation . . . . . 4
3.2 Factors of Interest . . . . . 4
3.3 Equating Procedures . . . . . 6
3.4 Evaluation Criteria . . . . . 7

**4 Results** **9**
4.1 Conditional Results . . . . . 9
4.1.1 CSEE . . . . . 9
4.1.2 CBias . . . . . 11
4.1.3 CRMSE . . . . . 12
4.2 Overall Results . . . . . 13
4.2.1 SEE . . . . . 14
4.2.2 AB . . . . . 14
4.2.3 RMSE . . . . . 15
4.3 Major Findings . . . . . 16

**5 Discussion** **18**

**6 References** **19**

# List of Tables

1 Descriptive Statistics for New Form ($X1$) and Old Form Raw Scores 6
2 Form Differences in P-values (New - Old) . . . . . 6
3 Equating Relationships for the Population after Few Chains (Test 3) . . . . . 8
4 Mean of Equated Scale Scores after 11 Equating over 1,000 replications . . . . . 14

# List of Figures

1 Population distributions for new ($X_1$) and old forms. . . . . 5
2 Equating chains. . . . . 7
3 Conditional SEE for Test 3. . . . . 10
4 Conditional bias for Test 3. . . . . 11
5 Conditional bias for Test 2. . . . . 12
6 Conditional RMSE for Test 3. . . . . 13
7 Overall SEE as a function of the number of equating chains. . . . 15
8 Overall absolute bias as a function of the number of equating chains. . . . . 16
9 Overall RMSE as a function of the number of equating chains. . 17

# Abstract

Every testing program has an equating linkage plan to maintain a common scale across multiple testing administrations and often in the case of large-scale testing programs, to maintain a common scale across international administrations and test takers. Equating error accumulates over a series of equating, and it is critical to monitor the amount of accumulative equating error to avoid scale drift. Also, the cubic-spline postsmoothing method is often used to reduce random error in equating. This study examines the long-term effects of various degrees of postsmoothing on accuracy of equated scale scores with an intention to identify a smoothing parameter that produces the least accumulative equating error.

Results of the current simulation study suggest that after two or three equating in the chain, a larger smoothing parameter provides more accurate equating results than a smaller smoothing parameter as it dramatically reduces random error. No significant interaction effect is seen between a smoothing parameter, and the test length and the score distribution. The study results clearly suggest that regardless of the difficulty level of a test form, test length, and shape of scores distributions, a larger smoothing parameter is preferred for multiple equating.

# 1 Introduction

Postsmoothing methods are often used to reduce random error in equating and to improve the accuracy of equating functions computed from unsmoothed data (ACT, 2017; College Board, 2017a). Of the postsmoothing methods that have appeared in the literature, the cubic spline postsmoothing (Kolen, 1984) has been widely used in practice. For instance, the cubic spline postsmoothing is considered as a smoothing technique for equipercentile equating for the SAT equating (College Board, 2017b). Previous studies have demonstrated the effectiveness of postsmoothing in reducing equating error when an appropriate smoothing parameter is applied (Colton, 1995; Hanson, Zeng, & Colton, 1994; Kim, 2014; Liu & Kolen, 2011). The use of postsmoothing methods, however, does not necessarily lead to a decrease in equating error because it can introduce systematic error (Kolen & Brennan, 2014). In particular, when a large degree (or a parameter) of postsmoothing is used, the increase in systematic error may exceed the decrease in random error, which will result in larger overall equating error. Systematic error in the context of equating can be conceived as the difference between the expected equating function estimated from a particular equating method and the population equating function. On the other hand, random error results from sampling variability and is often indexed by standard error of equating. The primary goal of postsmoothing is to maximize the decrease in the total equating error by producing a substantial reduction in random error sufficient enough to compensate for the increase in systematic error.

Zeng (1995) investigated the impacts of various degrees of postsmoothing on the accuracy of equipercentile equating under a random groups design. The study results showed that using a moderate degree of postsmoothing can improve equating accuracy by reducing substantial random error without allowing for excessive systematic error. Despite the popularity of postsmoothing, however, besides Zeng (1995), no studies have been conducted that compare smoothing parameters for the cubic-spline postsmoothing method.

When equating is conducted for a chain of forms given in administrations over time, equating error is likely to accumulate. Guo (2010) examined the accumulative equating error for linear equating methods and noted that as more equating are performed, more errors are accumulated. The accumulative equating error might cause an equating strain in which examinees with the same ability level receive different scale scores depending on a test form taken, and the position of the form in the equating series.

Liang et al. (2017) considered two IRT-based equating plans that are most widely used in current large-scale assessments (e.g., Educational Testing Service, 2009; Florida Department in Education in Harcourt, 2007): the item-pool equating plan and the year-to-year equating plan. The former utilizes a calibrated item pool to achieve score comparability over forms whereas the latter involves a single base form to which subsequent new forms are linked back in a chain such that scores from the successive new forms are put onto the base form scale. The year-to-year equating plan is known to be more cost- and time-

effective due to its simplicity (Liang et al., 2017). However, the effectiveness in implementation comes at the cost of potential scale drift due to cumulative equating errors resulting from a series of equating (Guo, 2010, Guo, Liu, Dorans, & Feigenbaum, 2011; Haberman & Dorans, 2009).

Likewise, equating error due to postsmoothing may accumulate if equating results are obtained through a chain of equating. However, little information exists yet with respect to the long-term effects of postsmoothing on equating errors, with a particular focus on the selection of a smoothing parameter. This study examines the long-term effects of various degrees of postsmoothing on accuracy of equated scale scores under the year-to-year equating plan. Data were obtained that had been used to produce actual equating functions for a large-scale testing program with US and international administrations. This equating data was used to establish data simulations that reflected actual empirical equatings from the testing program (discussed in the Data Generation section below). The resulting simulations made it possible to evaluate how equating error accumulates under realistic testing conditions, for specific degrees of postsmoothing, for shorter and longer test lengths, and for longer and shorter equating chains.

## 2 Cubic-Spline Postsmoothing

Cubic-spline postsmoothing method was introduced by Kolen (1984) in an effort to minimize the sampling error involved in equipencentile relationships. The resulting smoothed relationships are expected not to deviate too much from an original unsmoothed line while reducing random error. The cubic-spline function, $\hat{d}_Y$, for integer score $x_i$ is

$$\hat{d}_Y(x_i) = \alpha_{0i} + \alpha_{1i}(x - x_i) + \alpha_{2i}(x - x_i)^2 + \alpha_{3i}(x - x_i)^3, \tag{1}$$

where $x$ is a non-integer score that is larger than or equal to $x_i$ but smaller than $x_i + 1$. The function is found such that the following criterion is satisfied

$$\frac{1}{x_{high} - x_{low} + 1} \sum_{i:low}^{high} \left[ \frac{\hat{d}_Y(x_i) - \hat{e}_Y(x_i)}{\hat{\sigma}\left[\hat{e}_Y(x_i)\right]} \right]^2 < S, \tag{2}$$

where $x_{high}$ and $x_{low}$ are the highest and lowest integer scores, respectively, in the smoothing range; $\hat{e}_Y(x_i)$ represents the equating function that converts Form $X$ score, $x_i$, to Form $Y$ score; $\hat{e}_Y(x_i)$ is the estimated standard error of equating; and $S$ denotes the postsmoothing parameter. Equation 2 implies that the degree of smoothing is controlled through a value of $S$, and a larger $S$ value allows the smoothed function to depart further from the original unsmoothed equating function. In practice, the postsmoothing method is applied only to a restricted score range to avoid undesirable impacts of extreme scores with only few examinees on the final equating function. Kolen (1984) recommended having the smoothing score range between .5 and 99.5 percentile ranks. The equating relationship outside this range is often found through linear interpolation.

One important feature of the cubic-spline postsmoothing is that the symmetry requirement of equating is approximated by averaging two spline functions: a function for putting Form $X$ scores onto a Form $Y$ scale and an inverse function for placing Form $Y$ scores onto a Form $X$ scale. The final equating relationship is determined as an average of the two spline functions.

Unlike postsmoothing in which the estimated equating function is smoothed after equating, presmoothing smooths score distributions prior to equating. Among various presmoothing techniques, log-linear presmoothing methods have been used and researched extensively (Kolen & Brennan, 2014). In the log-linear presmoothing methods, the use of a log-linear model allows for goodness-of-fit statistical significance tests in determining a smoothing parameter. Unlike the log-linear presmoothing methods, however, no universally accepted statistical test exists for the cubic-spline postsmoothing method that helps to select an optimal smoothing parameter. As a result, the choice of an $S$ value heavily depends on psychometricians' subjective judgements. There are several strategies, however, that can aid such decision (Kolen & Brennan, 2014).

Inspection of graphs has been widely recognized as a tool in choosing a degree of smoothing. Specifically, one might visually compare unsmoothed and smoothed equating functions for various values of $S$ and identify the $S$ value such that the resulting function does not deviate too much from the unsmoothed function remaining within the specified standard error band, while being smooth enough to reduce random error. Recently, Kim, Brennan, and Lee (2020) proposed a new statistic for evaluating the fitness of the cubic-spline postsmoothing method that takes various factors into account such as a standard error band, a degree of departure from unsmoothed one, and the degree of smoothness, but the statistic is limited in that no statistical testing is available and it still heavily relies on the information from fitted plots.

Another approach to determining a smoothing parameter is to examine the discrepancies in the central moments such as the mean and standard deviation between the equated score distribution and the Form $Y$ distribution. The smoothing parameter that yields the closest moments to those of Form $Y$ is generally preferred. Such decisions, however, are also subject to evaluators' judgment as there will usually be no unique smoothing parameter that generates the closest values with respect to all the moments examined, which makes the practitioners still left with a final subjective decision. Moreover, the impact of choosing a particular smoothing parameter has not been systematically examined in the literature, particularly in the context of a chain of equating. This study is intended to fill the gap in the existing body of the literature. As mentioned previously, the cubic-spline postsmoothing method is employed in many large-scale testing programs with US and non-US administrations, such as SAT (College Board, 2017a) and ACT (ACT, 2017). As such, it is imperative to understand the behavior of a smoothing parameter and its impact on the long-term equating accuracy.

# 3 Method

## 3.1 Data Generation

Five tests from a large-scale assessment with US and international administrations were used in this study, three of which had lengths typical of full tests, and two of which had shorter lengths typical of subscores. These five tests differ in terms of contents, measuring different knowledge and skills. For each test, there were eleven new forms administered in the random groups design (Kolen & Brennan, 2014). In the operational setting, all new forms were equated directly to an old form that was administered with the new forms, and this old form has a scale score conversion table. For the current study, the equating data were reorganized to mimic a year-to-year equating plan where each new form was treated as if it were administered solely each year and equated to its previous adjacent new form. When multiple new forms existed, equating continued until an equating relationship was found between the target new form and old form. In this study, final equating result was a raw-to-scale score conversion for the target new form, similar to the data and equatings produced operationally but with equatings performed to evaluate equating chains and the accumulative error resulting from those chains.

A quick inspection of the original data revealed that there were some score points with zero frequency, which would probably not occur in the population distribution. To resolve this problem, log-linear models were fitted to the original data, with parameters chosen such that the Akaike Information Criterion (AIC) statistic is minimized. Finally, the population distribution was found for each test form as an average of the fitted and actual distributions so that the resulting population distributions would closely resemble the original data, have nonzero frequencies at all score points, and not completely correspond to any model or smoothing method. Samples of 4,000 (close to the actual sample size) were drawn from the population distributions for 1,000 replications. Note that the samples were drawn for each equating chain independently in order to avoid underestimation of equating error resulting from dependency between one chain to the next.

## 3.2 Factors of Interest

Eleven levels of the number of equating were considered in this study using eleven new forms and one old form. As the main focus of the current study was to investigate the effect of accumulative equating error, the major study factor (the number of equating) was designed to have as many levels as possible given the data available. The maximum number of equatings was fixed to 11 to reflect the reality that testing programs may rescale their tests on a regular basis to avoid a scale drift caused by various conditions including changes in norm groups or in test content.

In addition to a varied number of equating in a chain, two more simulation factors were considered including the degree of postsmoothing and the scale

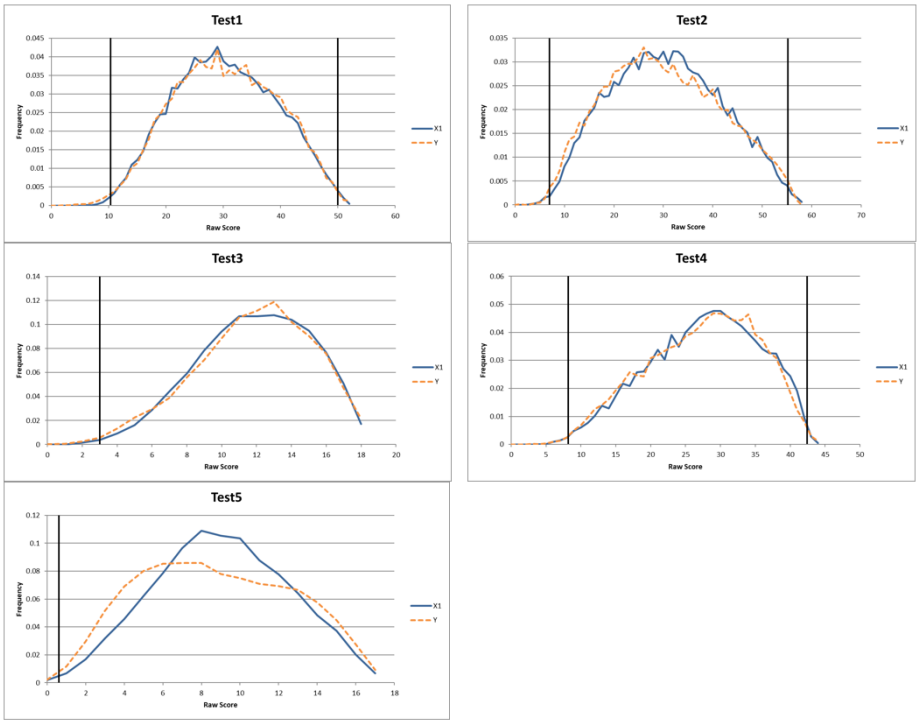


Figure 1: Population distributions for new ($X_1$) and old forms.

score range. First, eight levels of postsmoothing parameters were investigated: $S = 0.00$ (unsmoothed), 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. These particular parameters were chosen because of their frequent use in operational settings (SAT; College Board, 2017a) and those are the values suggested by Kolen and Brennan (2014). Also, an extensive review of the current literature revealed that a choice of postsmoothing parameters has been generally made within the range of [.01, .50] in research contexts (Hanson, Zeng, & Colton, 1994; Moses & Liu, 2011). Also, previous studies have demonstrated that the smoothing parameter within this range fits many large scale-assessment data (e.g., $S = .3$ in Cho, 2007; $S = .1$ in Lee, Lee, & Brennan, 2012).

Also, five levels of test length were examined: 58 (Test 2), 52 (Test 1), 44 (Test 4), 18 (Test 3), and 17 (Test 5). A raw-to-scale score conversion was arbitrarily created such that scale scores had the same score range with raw scores, except for Test 2, which has a scale score range of 200 - 800. The inclusion of Test 2 with a different scale range was intended to examine potential impact of the much wider score range on the study results. Finally, the shape of score distributions was also examined. A direct use of a set of real data allowed to work with various score distributions that are likely to be found with typical achievement tests. In Figure 1, distributions for new form 1 ($X_1$) and old form

Table 1: Descriptive Statistics for New Form ($X1$) and Old Form Raw Scores

| | Test1 | | Test2 | | Test3 | | Test4 | | Test5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Form | $X_1$ | $Y$ | $X_1$ | $Y$ | $X_1$ | $Y$ | $X_1$ | $Y$ | $X_1$ | $Y$ |
| Mean | 30.314 | 30.380 | 30.553 | 30.071 | 11.866 | 11.822 | 27.644 | 27.340 | 9.033 | 8.706 |
| S.D. | 8.927 | 9.094 | 11.146 | 11.573 | 3.255 | 3.356 | 7.910 | 7.975 | 3.468 | 3.923 |
| Skew. | .055 | -.009 | .111 | .164 | -.302 | -.402 | -.277 | -.302 | -.001 | .095 |
| Kurt. | -.707 | -.683 | -.732 | -.778 | -.488 | -.305 | -.623 | -.662 | -.564 | -.919 |

Table 2: Form Differences in P-values (New - Old)

| Test | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | Aver. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test1 | -.001 | -.020 | .006 | -.009 | -.003 | -.020 | .034 | -.028 | .002 | -.011 | -.006 | -.005 |
| Test2 | -.001 | .019 | -.002 | .020 | -.020 | .003 | -.004 | .002 | .020 | -.004 | .033 | .006 |
| Test3 | .002 | -.024 | -.015 | -.040 | -.007 | -.018 | .007 | -.032 | -.018 | -.040 | -.010 | -.018 |
| Test4 | .007 | -.053 | -.016 | -.029 | -.022 | -.057 | -.007 | -.058 | -.043 | -.024 | -.046 | -.032 |
| Test5 | .019 | .047 | .069 | .024 | .108 | .064 | .082 | .027 | .103 | .058 | .069 | .061 |

($Y$) are displayed for the five tests. Due to the large number of forms used in this study, the distributions for the other new forms were omitted. In general, as can be seen in Figure 1, the shape of distributions for the same test looked similar between the new and old forms. Tests 1 and 2 scores are positively skewed whereas Tests 3 and 4 have negatively skewed distributions. Last, Test 5 score are more symmetrical.

## 3.3 Equating Procedures

For each test, one new form ($X_1$) was chosen so that the new form has the closest average p-value (i.e., average raw score divided by the total number of items) to the old form ($Y$). The selection of the new form was made such that a fair comparison across the five tests could be achieved. Descriptive statistics for the target new form ($X_1$) and old form are summarized in Table 1. In Table 1, the new forms tend to be slightly easier than the old forms across tests, except for Test 1. The p-value differences between each new form and the old form can be found in Table 2. In Table 2, the last column represents the average p-value differences between the old form and each of the eleven new forms. In general, form differences for Tests 1 and 2 are small whereas form differences for Test 5 are relatively large.

In the year-to-year equating plan, scaling is conducted to establish a raw-to-scale score conversion for the base (or old) form ($Y$). Once the score scale is established, equating is performed to maintain the scale score equivalence for successive new forms through a chain of equating. To create such equating chain in this study, a varied number of new forms were added between $X_1$ and $Y$ as

$$
\begin{aligned}
&sc \leftarrow Y \leftarrow X_1 \\
&sc \leftarrow Y \leftarrow X_2 \leftarrow X_1 \\
&\vdots \\
&sc \leftarrow Y \leftarrow X_{11} \leftarrow X_{10} \leftarrow \cdots \leftarrow X_1
\end{aligned}
$$

Figure 2: Equating chains.

presented in Figure 2. For example, with a single equating in the chain, equating was performed only once from one new form ($X_1$) to the old form. Note that *sc* in Figure 2 stands for scale scores. With more new forms being added between $X_1$ and $Y$ that need to be equated, the number of equating increases while the initial form ($Y$) and the target form need to be equated ($X_1$) remain the same.

With 11 new forms, for instance, $X_{11}$ was equated to $Y$ first to obtain a scale-score conversion for $X_{11}$. Then, second equating was performed from $X_{10}$ to $X_{11}$ to find the conversion for $X_{10}$. The equating process continued until all the new forms in the chain were equated to the old form scale. As a result, the target new form ($X_1$) has a total of 11 conversion tables for the 11 conditions of the number of equating. The orders of eleven new forms were determined based on the p-values such that the resulting equating chain reflects realistic situations where the difficulty level of a test form changes unsystematically from one administration to the next.

Equipercentile equating with postsmoothing was conducted under the random groups design using Equating Recipes open source C functions (Brennan, Wang, Kim, & Seol, 2009). The final equating results were evaluated with respect to unrounded scale scores. Linear interpolation was applied for scores outside the percentile ranks of .5% to 99.5% for all replications.

## 3.4 Evaluation Criteria

The criterion equating relationship was defined for each test using the unsmoothed equipercentile equating for the single chain condition of $Y \leftarrow X_1$ based on the population distributions of $Y$ and $X_1$. Specifying the same form ($X_1$) as a target form to be equated, regardless of the number of equating, allows for the use of a single criterion equating relationship for all the equating chain conditions. Table 3 presents the criterion equating relationship for test 3, along with the target form's equating relationships based on one and two equating chains. Specifically, the second column of Table 3 shows the criterion equating relationship for the population distributions established based on a direct equating from $X_1$ to Y. The third and fourth columns in Table 3 show equating relationships of the same form $X_1$ for the population distributions but with one or two more added equatings between $X_1$ and $Y$. In this simulation setup, the criterion equating relationships were close to the operational equatings produced in the testing program, are empirical examples that can be evaluated on their

Table 3: Equating Relationships for the Population after Few Chains (Test 3)

| $X_1$ | $Y$ (Criterion) | $Y(\leftarrow X_2)$ | $Y(\leftarrow X_3 \leftarrow X_2)$ |
|---|---|---|---|
| 0 | .5000 | .5000 | .5000 |
| 1 | .5107 | .5112 | .5155 |
| 2 | .7120 | .7137 | .6965 |
| 3 | 1.2724 | 1.3698 | 1.3612 |
| 4 | 2.3877 | 2.3974 | 2.3419 |
| 5 | 3.4898 | 3.4644 | 3.4284 |
| 6 | 4.4946 | 4.4813 | 4.4767 |
| 7 | 5.5036 | 5.4993 | 5.5006 |
| 8 | 6.4523 | 6.4505 | 6.4499 |
| 9 | 7.3491 | 7.3470 | 7.3475 |
| 10 | 8.1984 | 8.1930 | 8.1904 |
| 11 | 8.9467 | 8.9457 | 8.9506 |
| 12 | 9.6319 | 9.6297 | 9.6264 |
| 13 | 10.2881 | 10.2920 | 10.2916 |
| 14 | 10.9895 | 10.9998 | 10.9984 |
| 15 | 11.7972 | 11.7971 | 11.7898 |
| 16 | 12.7429 | 12.7501 | 12.7012 |
| 17 | 13.8108 | 13.8087 | 13.7327 |
| 18 | 14.9749 | 14.9862 | 14.8800 |

own, and can also serve as criteria for evaluating error with respect to sampling, longer equating chains, and postsmoothing implementations.

Equating results were evaluated in terms of weighted standard errors of equating (SEE), bias, and root mean square error (RMSE), which represent the amount of random error, systematic error, and total error, respectively. Both overall and conditional results were investigated. Specifically, the conditional statistics were computed at each score point as follows:

$$CSEE(x) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left\{sc\left[\hat{e}_r(x)\right] - \left(\frac{1}{R}\sum_{r=1}^{R} sc\left[\hat{e}_r(x)\right]\right)\right\}^2}, \tag{3}$$

$$CBias(x) = \frac{1}{R}\sum_{r=1}^{R} sc\left[\hat{e}_r(x)\right] - sc\left[\hat{e}(x)\right], \tag{4}$$

and

$$CRMSE(x) = \sqrt{CSEE(x)^2 + CBias(x)^2}, \tag{5}$$

where $e(x)$ is the criterion equated score at x; $\hat{e}_r(x)$ is an estimated equated score at $x$ on $r^{th}$ replication; $R$ indicates the number of replications (i.e., $R$ =1,000),

and $sc[]$ denotes the conversion of the equated raw scores to the reporting scale. In order to assess the general performance of each smoothing parameter, overall statistics were also computed as:

$$SEE = \sum_x w_x CSEE(x), \tag{6}$$

$$AbsoluteBias(AB) = \sum_x w_x |CBias(x)|, \tag{7}$$

and

$$RMSE = \sum_x w_x \sqrt{CSEE(x)^2 + CBias(x)^2}, \tag{8}$$

where $w_x$ is the new group relative frequency at score $x$ in the population. Note that for the overall statistics, absolute bias (AB) was considered to measure the absolute magnitude of bias.

# 4 Results

Results are presented in both conditional (i.e., score level) and overall levels. The conditional results are presented first, followed by the overall results. Results for one test are mainly discussed in order to avoid redundancy due to the similar patterns observed across the five tests. Also, results for only a select set of smoothing parameters are displayed in plots for visual clarity (i.e., $S$ = .00, .01, .10, & .50). The last part of this section offers several major findings with respect to the study conditions of interest.

## 4.1 Conditional Results

Prior to conducting equating, score distributions for the new ($X_1$) and old forms were inspected as seen in Figure 1. Note that the vertical straight lines in Figure 1 indicate the lower and upper smoothing limits where linear interpolation is applied to obtain the equating relationship for scores outside this range. There is no upper limit associated with Tests 3 and 5 because the frequencies were more than .5% for all the score points, which mainly comes from the fact that those tests have relatively shorter test length (17 and 18). Given that the distributions provided in Figure 1 are based on the populations, sample distributions are unlikely to match the graphs exactly. Nonetheless, it is worth examining the shape of the score distribution to get a better insight into the conditional results that are presented next.

### 4.1.1 CSEE

CSEE for Test 3 can be found in Figure 3. The top left plot with a title of "Chain1" represents the CSEE results under a single equating chain (i.e., $Y \leftarrow X_1$). The next plot with a title of "Chain3" provides the results under three equating chains (i.e., $Y \leftarrow X_3 \leftarrow X_2 \leftarrow X_1$). For the sake of simplicity, results

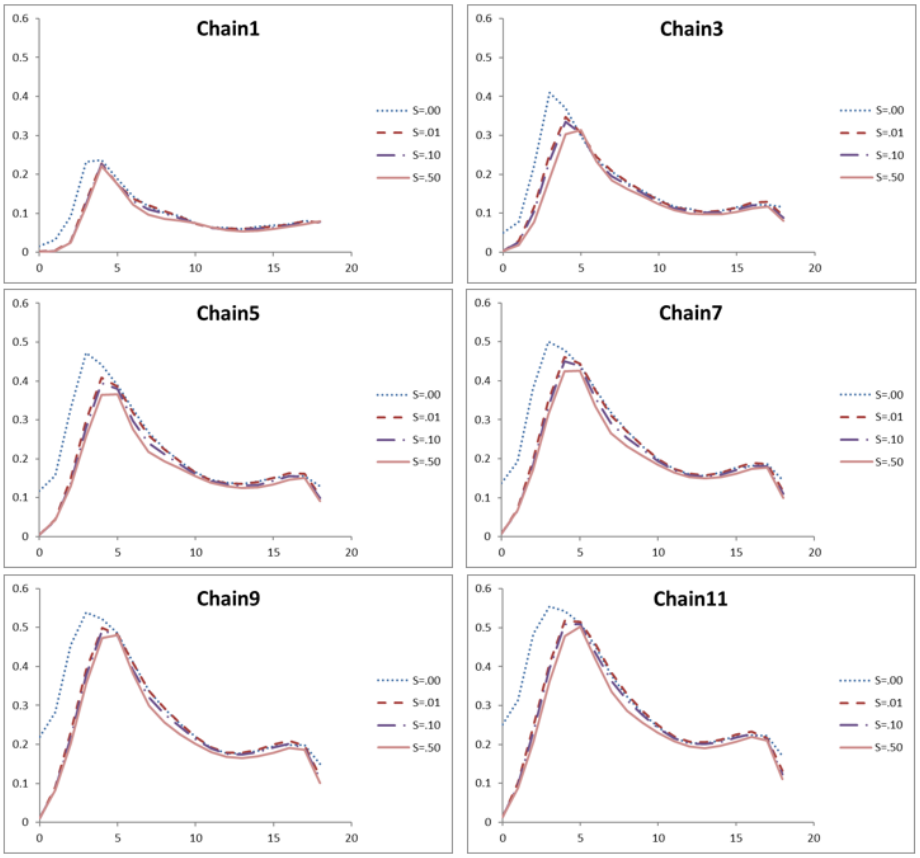


Figure 3: Conditional SEE for Test 3.

for an odd number of chains are presented to provide the overall picture with respect to the number of equating. In each plot, the vertical and horizontal axes represent CSEE and scale scores, respectively. The lines in the plots are the results for each of the smoothing parameter conditions from $S = 0.00$ to $S = 0.5$.

In Figure 3, it is apparent that CSEE is larger at the lower and upper ends of the score scale in which few examinees are located. Particularly, a substantial amount of CSEE is introduced at the lower score points where linear interpolation is applied. In this area, the unsmoothed equating tends to have a relatively larger CSEE than the other smoothing conditions. Another notable finding is that a smaller CSEE tends to be associated with a larger smoothing parameter, which is not surprising given that a larger $S$ value leads to a larger reduction in random error. As a result, the largest $S$ value of .5 produces the smallest CSEE across all score points among the smoothing parameters under investigation.

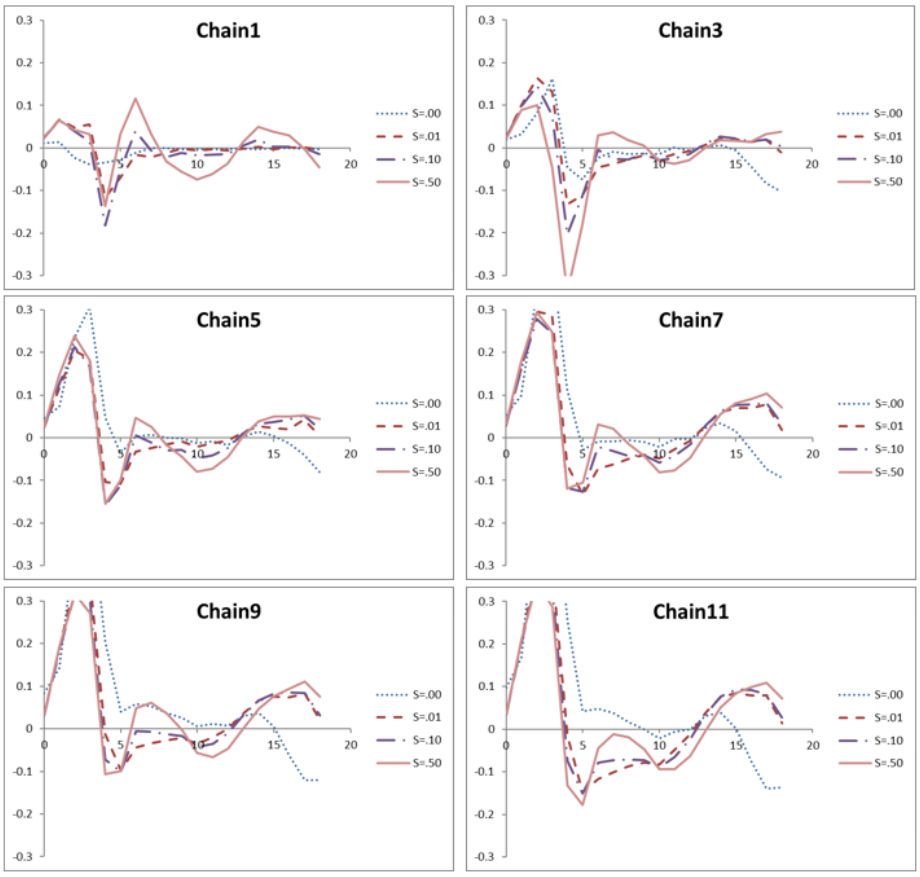


Figure 4: Conditional bias for Test 3.

### 4.1.2 CBias

CBias for Test 3 can be seen in Figure 4. Note that in Figure 4, the zero line serves as a baseline with no bias in equating results. Concerning the CBias for Test 3, larger smoothing parameters fluctuate more than smaller smoothing parameters, departing more from the baseline (i.e., zero bias). Particularly, a large amount of CBias is produced at the lower score points where linear interpolation is used.

It was seen that the patterns of CBias for tests with a fewer items are significantly different from those with more items. Results for Test 3 with 18 items (Figure 4) clearly reveals that the unsmoothed line behaves somewhat differently from the other conditions particularly for the lower and upper ends of the score range, which was not shown with Test 2 with 58 items (Figure 5). This tendency was found only with Tests 3 and 5 which had fewer items than the others. In order to offer a plausible explanation for this observation, one

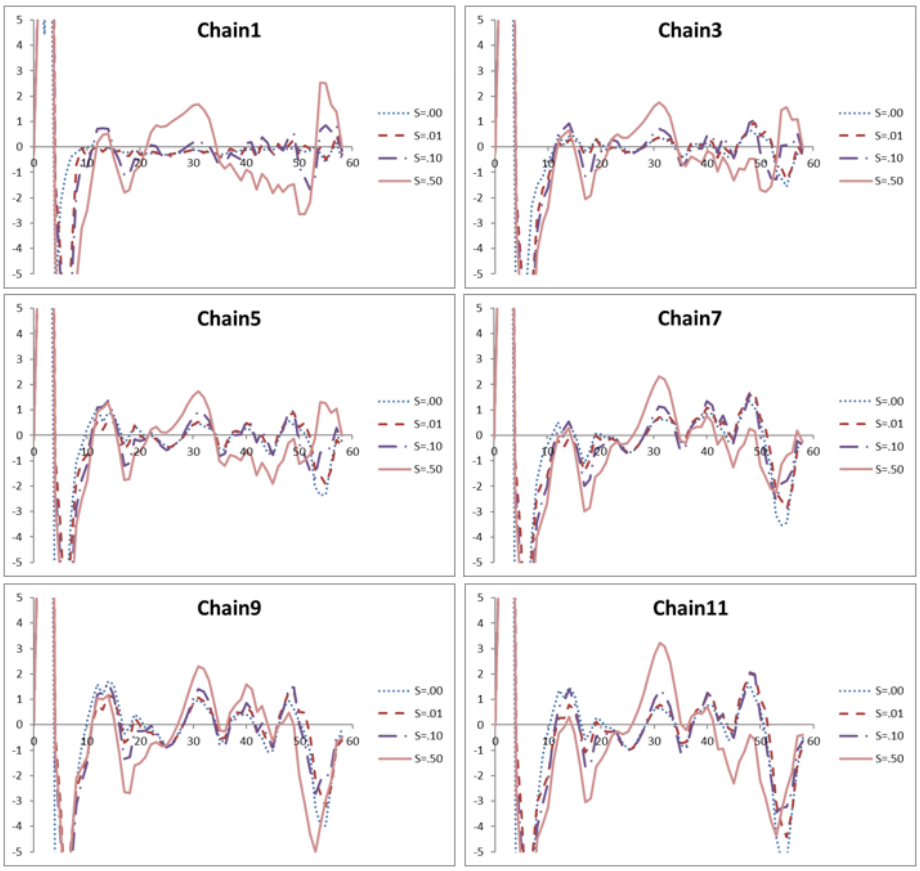


Figure 5: Conditional bias for Test 2.

replication of Test 3 under four equating chains was randomly selected and investigated with various specifications of the linear interpolation range and using one-directional smoothing ($Y \leftarrow X$ or $X \leftarrow Y$). From this simple experiment, it was apparent that the range of linear interpolation played an important role in causing discrepancies between the unsmoothed and smoothed equating results at the upper and lower ends of the score range. However, further research is needed to better explain why this peculiar pattern was seen for those with fewer items.

### 4.1.3 CRMSE

CRMSE for Test 3 is provided in Figure 6. In general, the overall pattern closely follows that of SEE. As with CSEE, a large amount of RMSE is observed at the lower end of scale scores, mainly due to the large amount of CSEE near the score range. It seems clear that the impact of CBias is less significant relative to

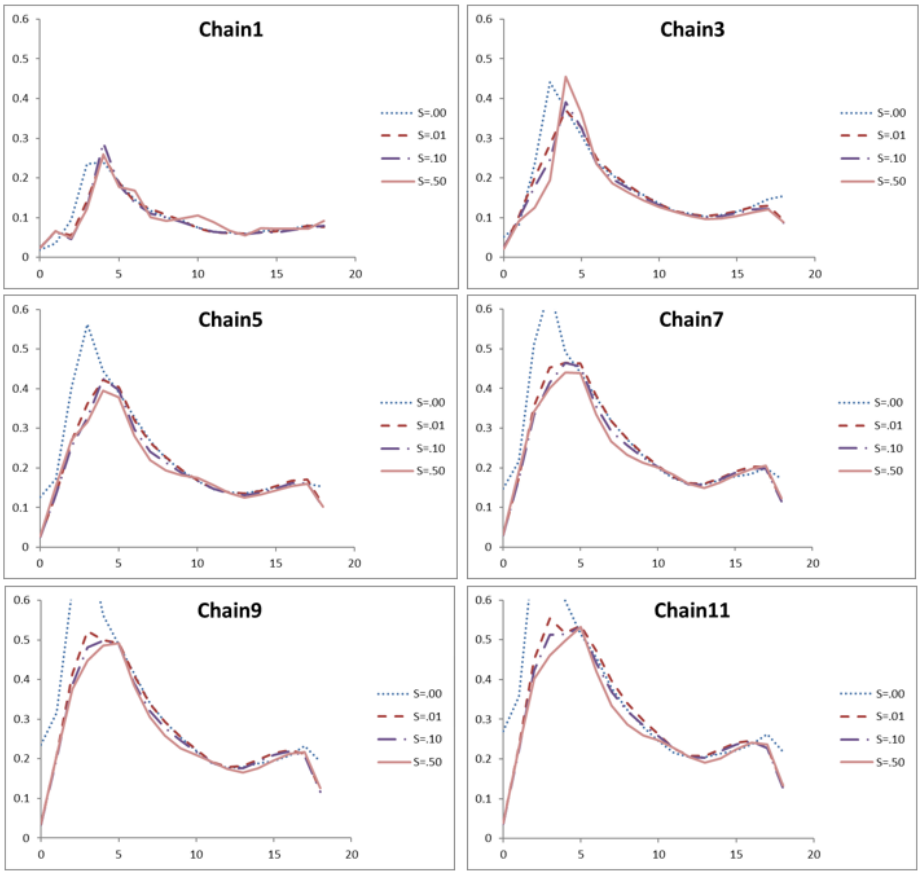


Figure 6: Conditional RMSE for Test 3.

that of CSEE, and consequently, larger smoothing parameters lead to a smaller total error (CRMSE) across all scale scores.

## 4.2 Overall Results

As noted earlier, $S$ values can also be chosen by inspecting the central moments. Table 4 summarizes the mean of the first moment (i.e., the mean) of equated scale scores over 1,000 replications for the condition of the number of equating chains equal to 11. The rightmost column presents the mean of the old form scale scores for each test. Note that in Table 4 the mean that is closest to the old form mean is in boldface type. One interesting observation is that for the tests with large p-value differences, a larger $S$ parameter tends to result in the mean closest to the old form mean. For example, the unsmoothed method produces the closest mean to the old form mean for Test 1 which has an average p-value difference of -.005, whereas $S$ of .5 yields the mean that is closest to the

Table 4: Mean of Equated Scale Scores after 11 Equating over 1,000 replications

| Test | $S = .00$ | $S = .01$ | $S = .05$ | $S = .10$ | $S = .20$ | $S = .30$ | $S = .40$ | $S = .50$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|
| Test1 | **27.641** | 27.647 | 27.648 | 27.648 | 27.649 | 27.649 | 27.648 | 27.647 | 27.599 |
| Test2 | 539.112 | **539.115** | 539.102 | 539.070 | 539.008 | 538.974 | 538.973 | 538.998 | 539.219 |
| Test3 | 9.473 | 9.470 | **9.470** | 9.470 | 9.470 | 9.468 | 9.467 | 9.465 | 9.470 |
| Test4 | 27.275 | 27.278 | 27.278 | 27.277 | **27.273** | 27.269 | 27.265 | 27.263 | 27.272 |
| Test5 | 8.967 | 8.957 | 8.958 | 8.959 | 8.963 | 8.968 | 8.974 | **8.979** | 8.979 |

*Note.* Bold numbers indicate the mean that is closed to the Form $Y$ mean.

old form mean for Test 5 which shows the largest difference of .061. According to the central moments criterion, in sum, it can be concluded that the larger the average difficulty differences between test forms, the larger $S$ value (e.g., $S = .5$) might be preferred.

#### 4.2.1 SEE

Figure 7 presents overall SEE results for the five tests. In general, larger SEE tends to be associated with smaller smoothing parameters. This pattern is consistently observed across the five tests. Also, a clear relationship is observed between the magnitude of SEE and the number of equating chains. That is, the more equating chains, the larger SEE regardless of the smoothing parameters used. Note that a different scale (i.e., y-axis) was used for Test 2 because of its larger scale score range relative to the other tests. It seems also apparent that the choice of score scales influences the magnitude of SEE: the longer the test, the larger SEE. As such, the use of a large smoothing parameter helps reduce more SEE in absolute magnitude for tests with more items. A comparison of Tests 1, 2, and 4 vs. Tests 3 and 5 clearly shows this tendency.

#### 4.2.2 AB

Results of absolute bias are provided in Figure 8. While SEE shows a monotonically increasing pattern as the number of equating chains increases, absolute bias does not necessarily follow the pattern, although there still is an overall trend that absolute bias grows as more equating is involved in the chain. Also, one tendency is clearly notable: a larger smoothing parameter tends to produce larger bias. It is important to note that the larger bias associated with Tests 4 and 5 relative to Test 1 is possibly due to the fact that the two tests have the largest form differences between the new and old forms as presented in Table 2. Also, the absolute amount of bias in equating is relatively small compared to that of SEE (see y-axes in Figure 7 and Figure 8).

With respect to absolute bias, Test 5 produces a somewhat unique pattern as seen in Figure 8. That is, unlike other tests, $S$ of .00 (i.e., no smoothing) adds more bias than other parameters as the number of equating increases. This unexpected pattern seems to be related to the shorter test length. The impact

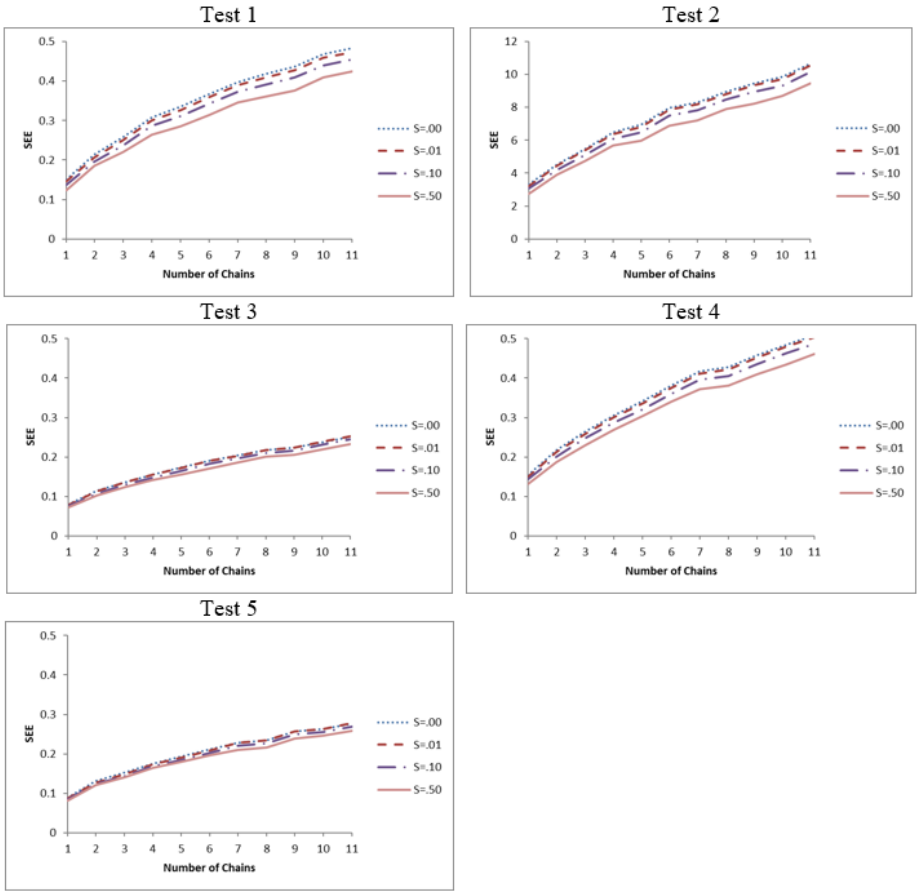


Figure 7: Overall SEE as a function of the number of equating chains.

of this distinguishing tendency for no smoothing seems more serious for Test 5 than Test 3. One possible reason for the relatively larger impact for Test 5 can be found from the shape of its score distributions as shown in Figure 1. Compared to Test 3, Test 5 has more symmetrical and flat distributions, which will likely be influenced greatly by extreme scores at both ends. Indeed, when examining the conditional results for the tests with a shorter test length (i.e., Tests 3 and 5), no-smoothing results deviated more from the zero bias line with a larger number of equating in the chain, particularly near both extreme ends.

### 4.2.3 RMSE

RMSE generally shows a monotonic increasing pattern with the added number of equating as presented in Figure 9. Most importantly, it is evident that a larger smoothing parameter results in smaller RMSE, primarily due to the greater

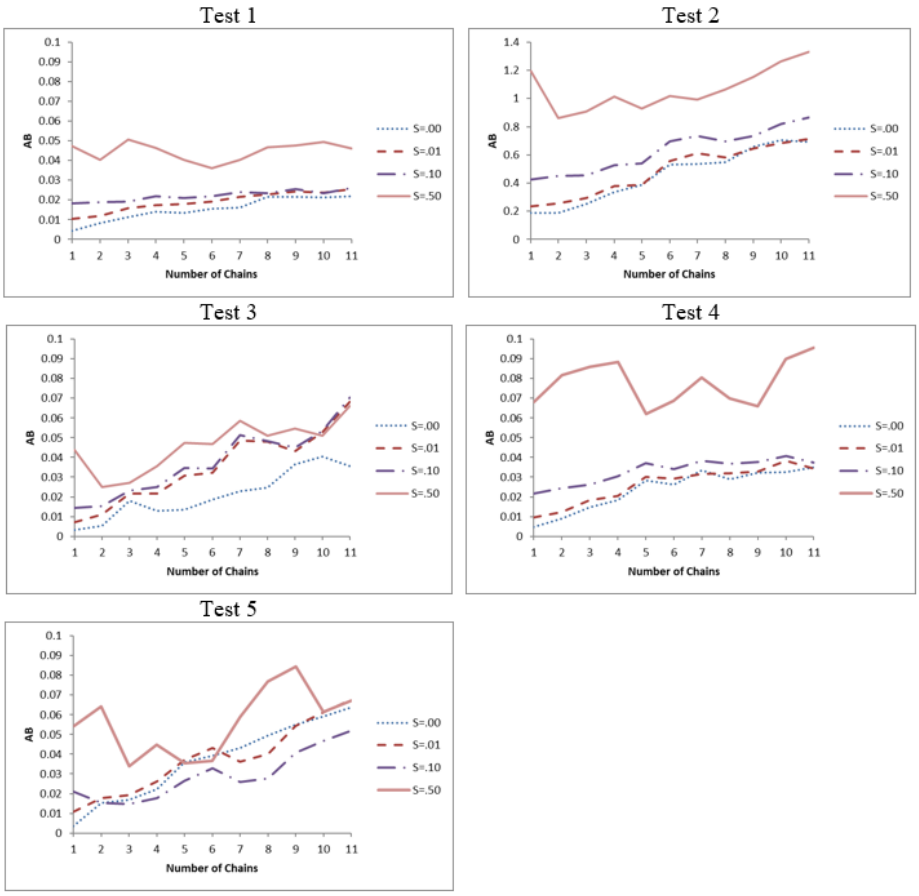


Figure 8: Overall absolute bias as a function of the number of equating chains.

impact of SEE than bias. The patterns of RMSE closely mirror those of SEE, as already observed with the conditional results. It is noteworthy that after the second or third equating, the largest $S$ value of .50 generates the least equating error regardless of test difficulty, test length, or the test distribution. Also, the relationship between the amount of equating error and the number of equating looks roughly linear, suggesting that one might be able to predict the degree of equating error after a few chains of equating even before administration.

## 4.3 Major Findings

The following provides a summary of the primary findings from the current study regarding the main study factors.

- The number of equating in the chain

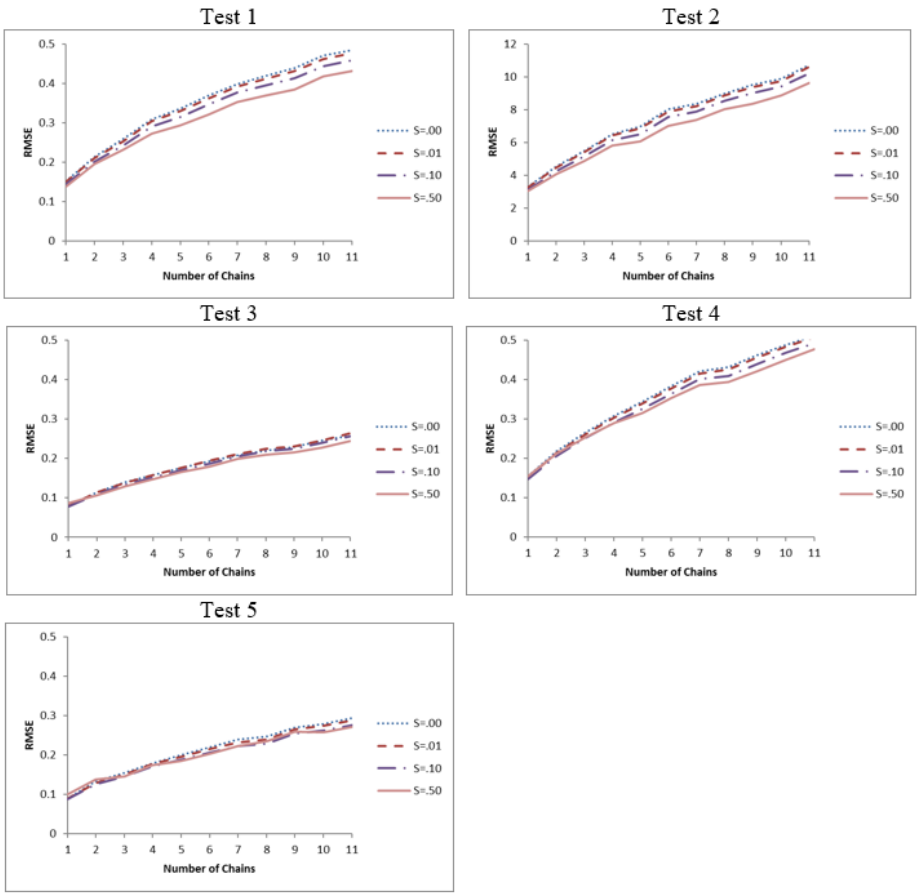


Figure 9: Overall RMSE as a function of the number of equating chains.

  - The larger the number of equating, the larger overall equating error.
  - The larger the smoothing parameter, the less overall equating error after two or three equating.

- Test length/shape of a score distribution
  - The longer the test is, the larger overall equating error in terms of absolute amount.
  - Regardless of the test length and the shape of a score distribution, the larger the smoothing parameter, the less equating error after second or third equating.

# 5 Discussion

Every testing program has an equating linkage plan to maintain a common scale across multiple testing administrations. These equating plans are especially important for maintaining the scale(s) of a large-scale testing program with US and international test takers in its testing population. For the series of equatings involved in the equating plan, equating error accumulates, and it is critical to monitor the amount of accumulative equating error to avoid scale drift. Liang et al. (2017) reported from their extensive review of the current public assessment technical reports for the 50 states in the U.S. between 2009 and 2016 that the year-to-year plan has been adopted by approximately 32% of the states for their standard state assessment programs. Although the year-to-year plan typically utilizes the common-item nonequivalent groups design, some testing programs consider the plan under the random groups design due to its efficiency and simplicity of implementation (e.g., SAT; College Board, 2017b). Despite its popularity in practice, however, only a limited number of research studies has been carried out with regard to which smoothing parameter leads to less accumulative equating error. This study compares various postsmoothing parameters under the year-to-year equating plan with an intention to identify a smoothing parameter that produces the least accumulative equating error. The results from this study can provide practical guidelines for the practitioners.

The results of the current study led to the following conclusions. First, after two or three equatings in the chain, a larger smoothing parameter provides more accurate equating results than a smaller smoothing parameter as it dramatically reduces random error. From this observation, a large smoothing parameter such as .5 or even larger is suggested when an equating plan includes multiple chained equating. Second, SEE increases as more equating is performed whereas changes in bias are less clear and systematic. Third, total equating error increases with the number of equating increases. Increased accumulative equating error is unavoidable with added equating, although using a large smoothing parameter somehow helps reduce total equating error. Oftentimes, rescaling is an option after several years of equating in order to avoid the scale drift that results from accumulative equating error. Last, no significant interaction effect is seen between a smoothing parameter, and test length and the score distribution. The study results clearly suggest that regardless of the difficulty level of a test form, test length, and shape of scores distributions, a larger smoothing parameter is preferred for multiple equating.

As is true for any studies, this study is not without limitations. First, the current study was limited by the fact that it did not consider various levels of form differences. More specifically, the order of new forms in the equating chain was fixed to be random in terms of form difficulty levels. In reality, however, it might be possible to have a new form with increasing or decreasing difficulty level in an equating chain. It might be interesting to see whether the similar results observed in this study are found under a different order of new form difficulty level. Another limitation is that this study was conducted using a fixed sample size. Although the sample size was determined based on the

practical experiences, fixing the condition makes it difficult to generalize the study results into other settings with a different sample size. Third, bias statistics were computed based on the criterion equating relationships established through unsmoothed equipercentile equating with the population data. However, no *perfect* equation criterion exists in any simulation study, and thus, bias statistics reported in this paper may not have fully captured the actual amount of bias that exists in reality. Further, if bias were considered as a primary error statistic, a completely different conclusion would have been obtained. The current research findings were largely drawn as a function of SEE, but less of bias due to its strong dominance in RMSE. In practice, it is also important to consider the amount of bias even though it is relatively smaller than the SEE. Future research might use multiple criterion such as a Difference That Matters to identify an acceptable level of (conditional) bias along the score scale. Despite these limitations, this study provides psychometricians with the implications of long-term postsmoothing equipercentile equating under various postsmoothing parameters.

## Acknowledgement

The authors appreciate the comments of Won-Chan Lee on the earlier draft of this paper.

# 6 References

ACT. (2017). *The ACT Technical Manual*. Iowa City, IA: ACT.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating Recipes (Version 1.0)* [Computer software]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Cho, Y. (2007). *Comparison of bootstrap errors of equating using IRT and equipercentile methods with polytomously-scored items under the common-item nonequivalent-groups design* (Unpublished doctoral dissertation). The University of Iowa, IA.

College Board. (2017a). *SAT Technical Manual: Characteristics of the SAT*. New York: The College Board.

College Board. (2017b). *SAT Suite of Assessments Technical Manual*. New York: The College Board.

Colton, D. A. (1995). *Comparing smoothing and equating methods using small sample sizes.* (Unpublished doctoral dissertation). The University of Iowa, IA.

Educational Testing Service. (2009). *California Standards Tests technical report spring 2008 administration*. Retrieved from: https://star.cde.ca.gov/techreports/CST/cst08techrpt.pdf

Harcourt. (2007). *Reading and mathematics technical report for 2006 FCAT test administrations*. San Antonio, TX: Harcourt Assessment. Retrieved from: https://docplayer.net/124212822-Technical-report-for-2006-fcat-test-administrations.html

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika, 75*, 438–453.

Guo, H., Liu, J., Dorans, N., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift*. Retrieved from https://www.ets.org/Media/Research/pdf/RR-11-46.pdf

Haberman, S., & Dorans, N. J. (2009). *Scale consistency, drift, stability: Definitions, distinctions and principles*. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, San Diego, CA.

Hanson, K., Zeng, L., & Colton, D. (1991, April). *A comparison of presmoothing and postsmothing methods in equipercentile equating* (ACT Research Report Vol. 94, No. 4). Iowa City, IA: American College Testing Program.

Kim, H. J., Brennan, R. L., & Lee, W. (2020). A new statistic to assess fitness of cubic-spline postsmoothing. *Journal of Educational Measurement, 57*, 124–144.

Kim, H. Y. (2014). *A comparison of smoothing methods for the common item nonequivalent groups design*. (Unpublished doctoral dissertation). The University of Iowa, IA.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Measurement, 9*, 25–44.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Lee, E., Lee, W., & Brennan, R. L. (2012). *Exploring equity properties in equating using AP Examinations* (Research Report No. 2012-4). New York: The College Board.

Liang, X., Koo, J., Yürekli, H., Paek, I., Becker, B. J., Binici, S., & Fukuhara, H. (2017). An empirical investigation of item-pool and year-to-year equating plans: using large-scale assessment data. *Florida Journal of Educational Research, 55*, 1–18.

Liu, C., & Kolen, M. J. (2011). Evaluating smoothing in equipercentile equating using fixed smoothing parameters. In M. J. Kolen, & W. Lee (Eds.), *Mixed format tests: Psychometric properties with a primary focus on equating (Volume 1)* (CASMA Monograph No. 2.1)(pp.213-236). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Moses, T., & Liu, J. (2011). *Smoothing and equating methods applied to different types of test score distributions and evaluated with respect to multiple equating criteria* (ETS Research Report Series, 2011-1). Princeton, NJ: Educational Testing Service.

Zeng, L. (1995). The optimal degree of smoothing in equipercentile equating with postsmoothing. *Applied Psychological Measurement, 19*, 177–190.