*Center for Advanced Studies in Measurement and Assessment*

*CASMA Research Report*

*Number 52*

# A Statistical Criterion to Assess Fitness of Cubic-Spline Postsmoothing

*Hyung Jin Kim*[†]
*Robert L. Brennan*
*Won-Chan Lee*

August, 2017
Revised June, 2019

[†]Hyung Jin Kim is Associate Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: hyungjin-kim@uiowa.edu). Robert L. Brennan is retired E. F. Lindquist Chair in Measurement and Testing and Founding Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: robert-brennan@uiowa.edu). Dr. Brennan is also a consultant to the College Board. Won-Chan Lee is Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: won-chan-lee@uiowa.edu).

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 A New Statistical Criterion** **1**
2.1 One Standard Error (1SE) Band . . . . . 2
2.2 Relative Deviation . . . . . 2
2.3 Relative Bumpiness . . . . . 3
2.4 Weighted Penalized Standardized Sum . . . . . 4

**3 Methodology** **6**
3.1 Applications to Real Datasets . . . . . 6
3.2 Simulation Study . . . . . 6
3.2.1 Evaluation . . . . . 7
3.2.2 Expected Results . . . . . 8

**4 Results** **9**
4.1 Applications to Real Datasets . . . . . 9
4.1.1 K & B Example . . . . . 9
4.1.2 Operational Dataset . . . . . 10
4.2 Simulation Study . . . . . 12
4.2.1 Overall Results . . . . . 12
4.2.2 Effect of Test Length . . . . . 14
4.2.3 Effect of Sample Size . . . . . 15

**5 Summary and Discussion** **16**

**References** **18**

# List of Tables


1 Descriptive Statistics for Real Datasets . . . . . 6
2 K & B Example: Weighted Penalized Standardized Sum of Squares for Various Pairs of Weights . . . . . 9
3 Operational Dataset: Weighted Penalized Standardized Sum of Squares for Various Pairs of Weights . . . . . 12
4 Weighted Overall Statistics . . . . . 13
5 Frequency Distributions of Postsmoothing Degree $S^*$ for Different Weights 13
6 Weighted Overall Statistics for Different Test Lengths Averaged over Sample Size Conditions . . . . . 14
7 Frequency Distributions of Postsmoothing Degree $S^*$ for Different Weights and Different Test Length Conditions . . . . . 15
8 Weighted Overall Statistics for Different Sample Sizes Averaged Over Test Length Conditions . . . . . 16
9 Frequency Distributions of Postsmoothing Degree $S^*$ for Different Weights and Sample Size Conditions . . . . . 17

# Abstract

In equating, smoothing techniques are frequently used to diminish sampling error. There are typically two types of smoothing: presmoothing and postsmoothing. Each smoothing technique has a smoothing degree that indicates how much smoothness is applied relative to observed frequency distributions or equipercentile relationships. For polynomial log-linear presmoothing, the choice of a lower-order polynomial degree (also known as a smoothing degree) can be determined statistically based on the Akaike information criterion or the Chi-square difference criterion. However, for cubic-spline postsmoothing, there is no current statistical criterion for choosing an optimum degree of smoothing. It is desired that the choice of smoothing degree results in a smooth function without departing too much from unsmoothed equivalents; currently, visual inspection has been an important tool in choosing the optimum degree, but visual inspection can be very subjective. Therefore, there is a need for development of a statistical criterion to choose an optimum smoothing degree. This study introduces a new statistical criterion for assessing the fitness of the cubic-spline postsmoothing method. The proposed statistical criterion accommodates three conditions: (1) one standard error band, (2) deviation from unsmoothed equivalents, and (3) smoothness, as discussed by Kolen and Brennan (2014). The proposed statistical criterion is certainly an improvement over visual inspection because the same weighting scheme for deviation and smoothness can be used for multiple equatings, whereas visual inspection cannot guarantee that the amount of attention given to deviation and smoothness are consistent across multiple equatings. Therefore, the authors suggest that this new statistical criterion may be helpful for assessing the fitness of postsmoothing.

# 1 Introduction

In equating, smoothing techniques are frequently used to diminish sampling error. There are typically two types of smoothing: presmoothing and postsmoothing. There are two main differences between presmoothing and postsmoothing: (1) when smoothing is performed and (2) the objects that are smoothed. For presmoothing, score distributions are smoothed prior to equating, whereas for postsmoothing, the equipercentile relationship is smoothed after equating. Each smoothing technique has a smoothing parameter that indicates how much smoothness is applied relative to the observed frequency distributions or the equipercentile relationships.

For polynomial log-linear presmoothing, the choice of a lower-order polynomial degree (also known as a smoothing degree) is determined statistically based on the Akaike information criterion ($AIC$) (Akaike, 1981) or the Chi-square difference criterion (Haberman, 1974). Note that, for the log-linear presmoothing method, a smaller lower-order polynomial degree implies more smoothing.

However, for cubic-spline postsmoothing, there is no current statistical criterion for choosing an optimum degree of smoothing. It is desired that the choice of smoothing degree results in (i) a smooth function that (ii) does not depart too much from unsmoothed equivalents. Visual inspection has been an important tool in choosing an optimum degree, but it can be very subjective; in other words, it is possible that the amount of attention given to (i) and (ii) varies from one equating to another. Therefore, it is important to develop a statistical criterion to choose an optimum degree. This study introduces a new statistical criterion for assessing "fitness" for the cubic-spline postsmoothing method. Considering that cubic-spline postsmoothing is used for some real testing programs (e.g., SAT (College Board, 2016) and ACT (ACT., 2014)), a statistical criterion for choosing an optimum postsmoothing degree will facilitate the process of selecting operational equating procedures.

For the rest of this paper, new and old forms are referred to as Form X and Form Y, respectively. Furthermore, throughout the paper, statistics are discussed in terms of differences between Form Y equivalents of Form X scores (i.e., $e_Y(x)$) and Form X scores (i.e., $x$), which will be referred to as "difference-scores," hereafter. The difference-score at Form X score $x_i$ can be mathematically expressed as $e_Y(x_i) - x_i$ and will be denoted as $de_Y(x_i)$. Note that Form Y equivalents of Form-X scores are used interchangeably with "equated-equivalents."

# 2 A New Statistical Criterion

According to Kolen and Brennan (2014), an optimum postsmoothing degree $S^*$ should result in *smoothed equated-equivalents* that *do not deviate too much from unsmoothed equated-equivalents*: a practical standard for deviation is that the smoothed relationship should lie mainly *within the one standard error (SE) band*. Consequently, there are three conditions to be examined to select the optimum postsmoothing degree: (1) one SE band, (2) deviation of smoothed equated-equivalents from unsmoothed equated-equivalents,

and (3) smoothness (bumpiness) of a smoothed function.

According to the deviation condition, the smaller deviation a smoothing degree gives, the more optimum the degree becomes. By contrast, the smoothness condition implies that a smoothing degree becomes more optimum as it yields larger smoothness over an equating relationship. As can be noticed, the relationship between degrees of deviation and optimum is opposite to the relationship between degrees of smoothness and optimum: a degree becomes more optimum as a numerical value representing degrees of deviation gets smaller, whereas a degree becomes more optimum as a numerical value representing degrees of smoothness gets larger. To avoid this potential confusion, the word "smoothness" will be replaced by "bumpiness" hereafter, with bumpiness viewed as a penalty for not being smooth.

Note that visual inspection typically considers difference-type plots in terms of difference-scores. Therefore, this study also uses difference-scores to examine equating results with respect to the three conditions [1].

## 2.1 One Standard Error (1SE) Band

For score $x_i$ across a range of interest, it is generally suggested that the cubic-spline postsmoothing method with an appropriate degree $S$ should have equated-equivalents within an one standard error (1SE) band around unsmoothed equated-equivalents. A range of interest $(x_{low}, x_{high})$ is often chosen to exclude score points with percentile ranks below 0.5 and above 99.5, where frequencies are generally very small. Therefore, the 1SE band condition excludes degrees that result in even one equated-equivalent outside 1SE bands in the range of interest. The 1SE band condition should be satisfied prior to considering the other two conditions, described next.

## 2.2 Relative Deviation

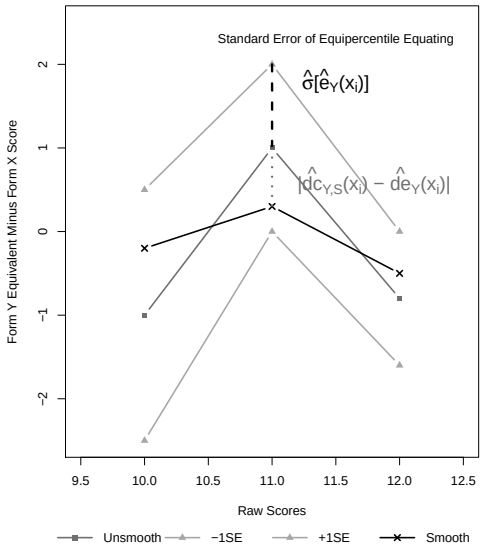
For the deviation condition, this study suggests that deviations of smoothed equivalents from unsmoothed equivalents should be measured relative to standard errors. Therefore, the study proposes that the relative deviation be measured by an average standardized squared difference between smoothed and unsmoothed equated-equivalents as follows:

$$D = \frac{\sum_{i=low}^{high} \left[ \frac{\hat{dc}_{Y,S}(x_i) - \hat{de}_Y(x_i)}{\hat{\sigma}[\hat{e}_Y(x_i)]} \right]^2}{high - low + 1}, \tag{1}$$

where $x_{low}$ and $x_{high}$ are the lower and upper integer scores in the range of interest; $\hat{de}_Y(x_i)$ is an equated-equivalent at score $x_i$ for the unsmoothed equipercentile equating relationship minus score $x_i$; $\hat{dc}_{Y,S}(x_i)$ is an equated-equivalent at score $x_i$ for the cubic-spline method with a degree $S$ minus score $x_i$; $\hat{\sigma}[\hat{e}_Y(x_i)]$ is the estimated standard

[1] The three criteria are started with respect to Form Y equivalents of Form X scores rather than difference-scores that are typically used operationally for postsmoothing. The use of difference-scores is discussed further in the Appendix

Figure 1: Illustration of Relative Deviation




error of equipercentile equating; and, $\hat{e}_Y(x_i)$ is an equated-equivalent at score $x_i$ for the unsmoothed equipercentile equating relationship.

The formula is standardized in a sense that a value for the squared bracket in Equation 1 is always between 0 and 1 for each $x_i$. Note that, because of the 1SE band condition, $|\hat{dc}_{Y,S}(x_i) - \hat{de}_Y(x_i)|$ is always less than $\hat{\sigma}[\hat{e}_Y(x_i)]$ (See Figure 1 for a graphical illustration). Consequently, the relative deviation as defined in Equation 1 is always between 0 and 1. With respect to the relative deviation, as $D$ declines, the associated smoothing degree becomes more optimum.

## 2.3 Relative Bumpiness

For the bumpiness condition, the study suggests that bumpiness of a smoothed function should be described in relation to bumpiness of an unsmoothed function. This study proposes that bumpiness of a smoothed function relative to an unsmoothed function be measured as follows:

$$B = \frac{\sum_{i=low+1}^{high-1} |\{\hat{dc}_{Y,S}(x_{i+1}) - \hat{dc}_{Y,S}(x_i)\} - \{\hat{dc}_{Y,S}(x_i) - \hat{dc}_{Y,S}(x_{i-1})\}|}{\sum_{i=low+1}^{high-1} |\{\hat{de}_Y(x_{i+1}) - \hat{de}_Y(x_i)\} - \{\hat{de}_Y(x_i) - \hat{de}_Y(x_{i-1})\}|}, \qquad (2)$$

where, for example, $|\hat{dc}_{Y,S}(x_i) - \hat{dc}_{Y,S}(x_{i-1})|$ is the magnitude of change for the difference in equated-equivalents for the cubic-spline method with a degree $S$ as $x$ increases from $x_{i-1}$ to $x_i$.

In order to examine bumpiness as defined in Equation (2), $\gamma_1$ is used to refer to the magnitude of change in difference-scores, $\hat{dc}_{Y,S}(x_i) - \hat{dc}_{Y,S}(x_{i-1})$, as $x$ increases from $x_{i-1}$ to $x_i$; and, $\gamma_2$ is used to refer to the magnitude of change in difference-scores,

$\hat{d}c_{Y,S}(x_{i+1}) - \hat{d}c_{Y,S}(x_i)$, as $x$ increases from $x_i$ to $x_{i+1}$. The numerator in Equation (2) then becomes $|\gamma_2 - \gamma_1|$, and it quantifies bumpiness at score $x_i$. If signs are different for $\gamma_1$ and $\gamma_2$, the bumpiness measure is equal to $|\gamma_1| + |\gamma_2|$. Figure 2 shows how the bumpiness at score $x_i$ can be understood graphically. For example, if bumpiness occurs as a concave-down pattern as illustrated in Figure 2(a) where $\gamma_1 > 0$ and $\gamma_2 < 0$ (i.e., signs are different), the quantified bumpiness at score $x_i$ is equal to $|\gamma_1| + |\gamma_2|$. Similarly, if bumpiness occurs as a concave up pattern as illustrated in Figure 2(b) where $\gamma_1 < 0$ and $\gamma_2 > 0$, the bumpiness measure is also equal to $|\gamma_1| + |\gamma_2|$.

However, if signs of $\gamma_1$ and $\gamma_2$ are the same (e.g., difference-scores at $x_{i-1}$, $x_i$, and $x_{i+1}$ tend to increase as depicted in Figure 2(c)), bumpiness is quantified as $|\gamma_1| - |\gamma_2|$ if $|\gamma_1| > |\gamma_2|$ or as $|\gamma_2| - |\gamma_1|$ if $|\gamma_2| > |\gamma_1|$. In doing so, for indistinct bumpiness at score $x_i$, a measure of bumpiness becomes smaller than that for distinct bumpiness. The same rule applies when difference-scores at $x_{i-1}$, $x_i$, and $x_{i+1}$ tend to decrease. For example, for the smoothing degree $S = 1.00$ where difference-scores form a close-to-linear line, the numerator in Equation (2) results in overall bumpiness close to zero. As a result, the bumpier the function is around $x_i$, the larger the changes in difference-scores as $x$ increases from $x_{i-1}$ to $x_i$ and from $x_i$ to $x_{i+1}$; and eventually, $|\{\hat{d}c_{Y,S}(x_i) - \hat{d}c_{Y,S}(x_{i-1})\} - \{\hat{d}c_{Y,S}(x_{i+1}) - \hat{d}c_{Y,S}(x_i)\}|$ becomes larger as the function gets bumpier. Therefore, the numerator and denominator in Equation (2) quantify the overall bumpiness of the cubic-spline postsmoothed equipercentile relationship with a smoothing degree $S$ and the overall bumpiness of the unsmoothed equipercentile relationship, respectively.

Since the smoothed equipercentile function is expected to have smaller up-and-downs than the unsmoothed function, relative bumpiness as defined in Equation (2) is also expected to be between 0 and 1. With respect to relative bumpiness, the smaller $B$ is, the less bumpy (i.e., smoother) an equating relationship is; and, the associated smoothing degree becomes more optimal.

### 2.4 Weighted Penalized Standardized Sum

Considering all three conditions, this study proposes consideration of a weighted penalized standardized sum of $D$ and $B$:

$$WP(S) = (w_D \times D) + (w_B \times B), \tag{3}$$

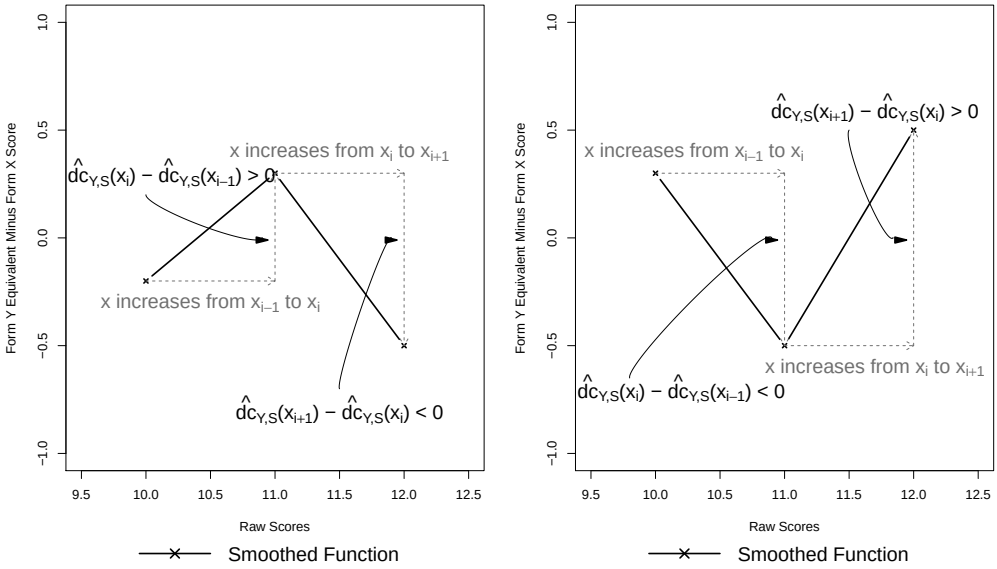
where $B$ can be viewed as a penalty function for not being smooth; $w_D + w_B = 1$; and $w_D, w_B \geq 0$. In doing so, $WP(S)$ always remains between 0 and 1. The authors propose that the optimum smoothing degree $S^*$ be the one that minimizes $WP(S)$, provided no smoothed value is outside the 1SE band. If a researcher wants to minimize the relative deviation of a smoothed relationship from an unsmoothed relationship, the optimum degree can be selected using $w_D = 1$ and $w_B = 0$. If the sole interest of a researcher is to have a relationship be as smooth as possible, $w_D = 0$ and $w_B = 1$ can be considered for selection of the optimum degree.

Weights, $w_D$ and $w_B$, can be any numbers as long as $w_D + w_B = 1$, and $w_D, w_B \geq 0$.

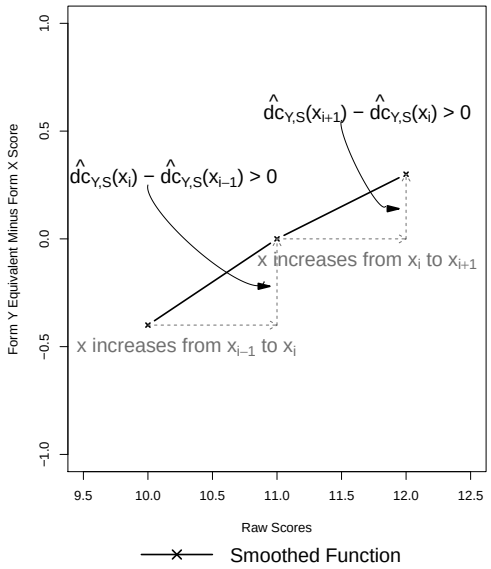
Figure 2: Illustration of Relative Bumpiness

(a) Concave Down

(b) Concave Up




(c) Linear




These weights can be determined based on researchers' interest or operational purposes. With a set of predetermined weights, researchers can examine the performance of smoothing degrees consistently across many equatings.

Table 1
*Descriptive Statistics for Real Datasets*

| | Test Form | Number of Examinees | Number of Items | Score Range (Min, Max) | $\hat{\mu}$ | $\hat{\sigma}$ |
|---|---|---|---|---|---|---|
| K & B | Form Y | 4,152 | 40 | (0, 40) | 18.9798 | 8.9393 |
| | Form X | 4,329 | 40 | (0, 40) | 19.8524 | 8.2116 |
| Operational | Form Y | 3,984 | 35 | (0, 35) | 21.0703 | 6.2017 |
| | Form X | 3,935 | 35 | (0, 35) | 19.3827 | 6.5873 |

# 3 Methodology

Unfortunately, there is currently no statistical criterion for choosing an optimum degree of cubic-spline postsmoothing, so there is no way to evaluate the accuracy of this newly proposed statistical criterion. Therefore, this study applies the statistics in Equations 1 to 3 to real datasets to examine whether or not decisions are sensible and congruent to those made based on visual inspection. In addition, a simulation study was conducted to observe whether or not the new statistical criterion gave results as expected for different study factors.

## 3.1 Applications to Real Datasets

For the first real dataset, this study employed the example that Kolen and Brennan (2014) used to present the application of cubic-spline postsmoothing where equating was conducted under the random groups design. For the second example, the study considered one operational dataset from a large-scale assessment. Table 1 provides summary of descriptive statistics for the real datasets.

For both datasets, cubic-spline postsmoothing was performed with linear interpolation to obtain equated-equivalents for percentile ranks below 0.5% and above 99.5%. In other words, the 1SE band condition was employed with respect to the range of $(x_{low}, x_{high})$ such that $x_{low}$ and $x_{high}$ are score points with percentile ranks below 0.5 and above 99.5, respectively. For smoothing degrees, eight values were considered: 0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.75, and 1.00. These eight values are the degrees considered for the example in Kolen and Brennan (2014) to present the application of cubic-spline postsmoothing. The optimum degrees $S^*$ were then selected based on the weighted penalized standardized sum using five different weights. The five weights included $(1, 0)$, $(2/3, 1/3)$, $(1/2, 1/2)$, $(1/3, 2/3)$, and $(0, 1)$ for $(w_D, w_B)$.

## 3.2 Simulation Study

In addition to the study factors considered for the real datasets (i.e., weights and smoothing degrees), two more factors were considered for the simulation study: test length and sample size. For test length, the study considered three conditions: 20, 40, and 60 items. For sample size, 500, 3000, and 6000 were considered for both new and old groups.

In order to construct simulated tests, the 3 parameter logistic (3PL) item response theory (IRT) model was used. Item parameter estimates (i.e., $\hat{a}$, $\hat{b}$, and $\hat{c}$) were adopted from one operational test from a large-scale assessment. Item parameters were carefully chosen so that means and standard deviations of $a$ and $b$ were matched as closely as possible between Form X and Form Y. Ability parameters were generated from the normal distribution with the mean 0 and the standard deviation 1. For each study condition (test length and sample size), item responses were generated using the 3PL model for the new group taking Form X and the old group taking Form Y. Scoring was conducted in terms of the number of correct items.

Each simulated dataset was used for conducting cubic-spline postsmoothed equating with thirteen smoothing degrees of 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.75, and 1.00. More number of smaller degrees are considered to observe changes in the statistics in more fine-grained. Note that equating was conducted under the random groups design. The optimum degrees $S^*$ were then selected based on the weighted penalized standardized sum using the five different weights considered for the real datasets. For each study condition, five hundred datasets were simulated to compute random errors and systematic errors in equating results.

### 3.2.1 Evaluation

In order to evaluate equating results for the simulation study, the IRT observed-score equating method was considered as the population equating relationship (i.e., criterion). Ability parameters for 100,000 examinees were generated from the normal distribution with the mean 0 and the standard deviation 1. With the pre-selected item parameters, response patterns were generated separately for test lengths of 20, 40, and 60 items. Since test forms were constructed using the IRT model, the IRT observed-score equating method was expected to results in a relationship that was smooth enough to be considered as a population relationship.

Equating results were evaluated in terms of overall amount of error over the entire score scale. Three weighted statistics were considered: weighted average root mean squared bias (WRMSB), weighted average standard error of equating (WSE), and weighted average root mean squared error (WRMSE) that were computed as follows:

$$\text{WRMSB} = \sqrt{\sum_i w_i \text{Bias}_i^2} \tag{4}$$

$$\text{WSE} = \sqrt{\sum_i w_i \text{CSE}_i^2} \tag{5}$$

$$\text{WRMSE} = \sqrt{\sum_i w_i \text{RMSE}_i^2}, \tag{6}$$

where $w_i$ is a relative frequency for score point $x_i$ for the new group used for the criterion equating relationship. In Equations 4 to 6, $\text{Bias}_i$, $\text{CSE}_i$, and $\text{RMSE}_i$ were computed as

below:

$$\text{Bias}_i = \frac{\sum_{j=1}^{500}(\hat{e}_{Y,S^*}(x_i) - e_Y(x_i))}{500} \tag{7}$$

$$\text{CSE}_i = \sqrt{\frac{\sum_{j=1}^{500}(\hat{e}_{Y,S^*}(x_i) - \bar{\hat{e}}_{Y,S^*}(x_i))^2}{500}} \tag{8}$$

$$\text{RMSE}_i = \sqrt{\text{Bias}_i^2 + \text{CSE}_i^2}, \tag{9}$$

where $i$ is a score point; $x_i$ is a raw score at score point $i$; $j$ is the $j^{th}$ replication; $e_Y(x_i)$ is Form Y equivalent of Form X score $x_i$ for the criterion relationship; $\hat{e}_{Y,S^*}(x_i)$ is Form Y equivalent of Form X score $x_i$ for a cubic-spline postsmoothed equating relationship with the optimum degree $S^*$ based on the $WP(S)$ under a study condition; and, $\bar{\hat{e}}_{Y,S^*}(x_i)$ is the average of Form Y equivalent of Form X score $x_i$ for a cubic-spline postsmoothed equating relationship with the optimum degree $S^*$ based on the $WP(S)$ under a study condition over one hundred replications. Note that the overall statistics were computed using equating results for the optimum degrees.

The study also examined frequency distributions of the optimum degree $S^*$ for different weights of $(w_D, w_B)$ under study conditions, which allowed for observation of the impact of weighting schemes in choosing $S^*$ and the change in impact as the study conditions varied.

### 3.2.2 Expected Results

For the simulation study, expected results are as follows:

- The WRMSB tends to decrease as $w_D$ increases, whereas the WSE tends to decrease as $w_B$ increases.
- When $w_D = 1$ and $w_B = 0$, the smallest $S$ values among the considered will be selected as the optimum degree most frequently.
- As interest shifts from reducing the relative deviation to reducing the relative bumpiness (i.e., $w_D \downarrow 0$ and $w_B \uparrow 1$), larger $S$ values will be selected as the optimum. When $w_D = 0$ and $w_B = 1$, the largest degree $S$ whose equating relationship remains within the 1SE band will be selected as the optimum degree.
- As test length increases, larger $S$ values will be selected less frequently as the optimum degree and smaller $S$ will be selected more frequently as the optimum degree because, as test length increases, observed frequency distributions and unsmoothed equating relationships are bumpier, and it is easier for larger smoothing degrees to result in equated-equivalents outside the 1SE band.
- For a fixed test length, as sample size decreases, smaller $S$ values will be selected more frequently as the optimum degrees because, as sample size decreases, observed

Table 2
*K & B Example: Weighted Penalized Standardized Sum of Squares for Various Pairs of Weights*

| $(w_D, w_B)$ | S=0.01 | S=0.05 | S=0.10 | S=0.20 | S=0.30 | S=0.50 | S=0.75 | S=1.00 |
|---|---|---|---|---|---|---|---|---|
| (1, 0) | **0.0154** | 0.0638 | NA | NA | NA | NA | NA | NA |
| (2/3, 1/3) | 0.1251 | **0.1041** | NA | NA | NA | NA | NA | NA |
| (1/2, 1/2) | 0.1800 | **0.1243** | NA | NA | NA | NA | NA | NA |
| (1/3, 2/3) | 0.2349 | **0.1445** | NA | NA | NA | NA | NA | NA |
| (0, 1) | 0.3447 | **0.1848** | NA | NA | NA | NA | NA | NA |

*Note*: NA implies that there is at least one score outside the 1SE band. For each pair of weights, a boldfaced number refers the minimum $WP(S)$ among different $S$s excluding those for NA.

frequency distributions and unsmoothed equating relationship become bumpier. Conversely, as sample size increases, smaller $S$ values will be selected less frequently as the optimum degree. Simultaneously, as sample size increases, larger values will be selected less frequently as the optimum degree because of the 1SE band criterion. Note that standard errors become smaller for larger sample sizes.

# 4 Results

This section consists of two subsections. The first subsection presents results of the newly proposed statistical criterion applied to the real datasets. The second subsection shows results for the simulation study.

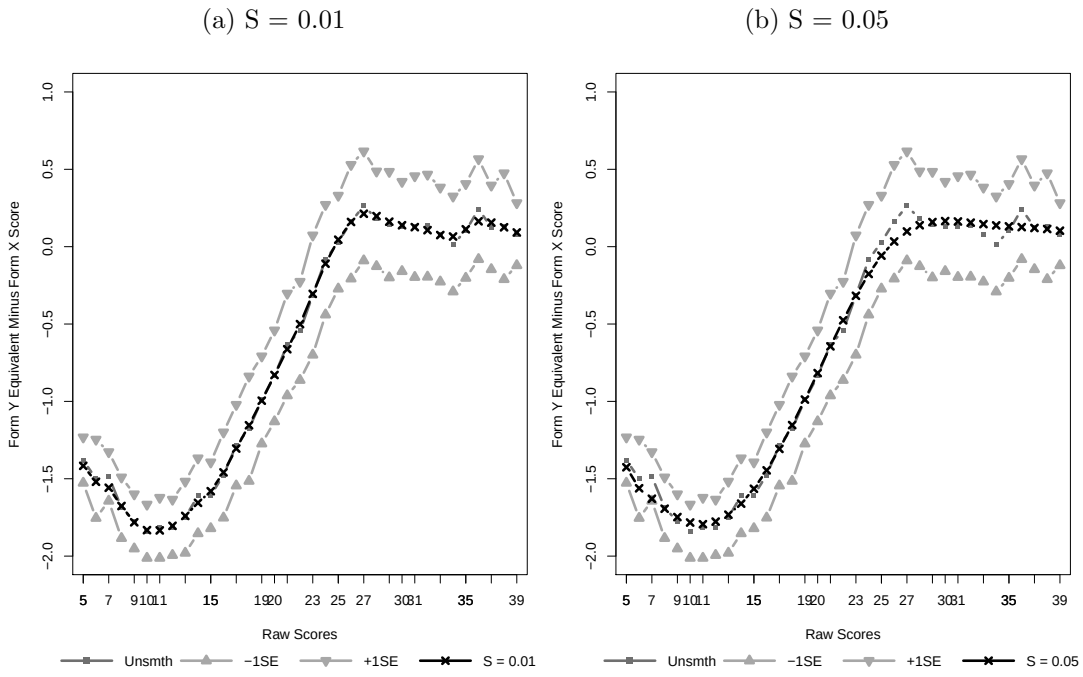
## 4.1 Applications to Real Datasets

This section consists of two subsections. The first subsection presents results for the example discussed in Kolen and Brennan (2014), and the second subsection presents results for applying the new statistical criterion to the operational dataset from a large-scale assessment.

### 4.1.1 K & B Example

This study took the example given in Kolen and Brennan (2014) (i.e., K & B example) and applied the new statistics to assess the fitness of cubic-spline postsmoothing. With respect to the 1SE band condition, postsmoothing degrees $S \geq 0.10$ had at least one raw score whose equated-equivalent was outside the 1SE band. Thus, the remaining two degrees (i.e., $S = 0.01$ and $S = 0.05$) were considered for further inspection with respect to the relative deviation and the relative bumpiness. Final weighted penalized standardized sum of squares was then computed for each of the five weights. Table 2 presents the statistics for the eight different postsmoothing degrees and the five different weights for $(w_D, w_B)$, where "NA" indicates that those smoothing degrees did not satisfy the 1SE band condition.

Figure 3: K & B Example: Raw-to-raw equivalents for postsmoothing $S = 0.01$ and 0.05




Based on Table 2, if the sole interest was to reduce the relative deviation (i.e., $w_D = 1$), $S = 0.01$ is the optimum degree. For the other choices of weights, $S = 0.05$ was selected as the optimum degree. The difference plot in Figure 3(a) shows that the equated-equivalents using $S = 0.01$ are quite bumpy at scores above 25, whereas the bumpiness disappeared for $S = 0.05$ as shown in Figure 3(b). For this example, $S = 0.05$ is the optimum choice as long as reducing bumpiness is a concern. For this example, it is quite clear which degree was the optimum because there were only two candidates.

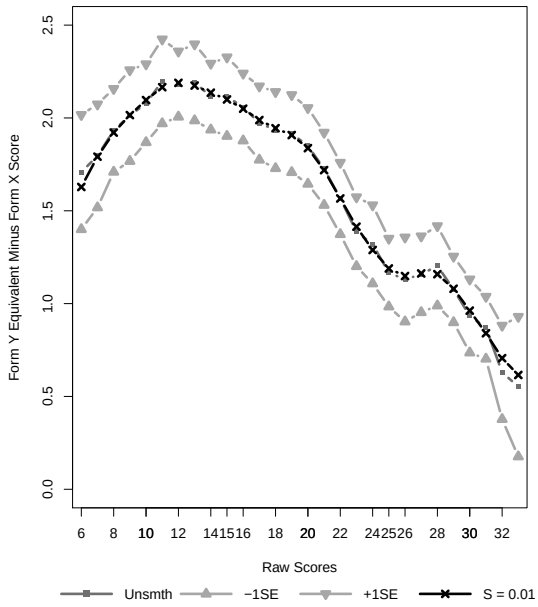
### 4.1.2 Operational Dataset

This study considered one operational data set from a large-scale assessment and obtained $WP(S)$ for the eight postsmoothing degrees and the five pairs of weights $(w_D, w_B)$. For the operational data set, this study excluded score points with percentile ranks below 0.5 and above 99.5. Consequently, the study considered smoothed equated-equivalents for raw scores ranging from 6 to 32 to examine whether or not the equivalents were outside the 1SE band condition. Postsmoothing degrees of $S \geq 0.30$ had at least one raw score whose equated-equivalent was outside the 1SE band. Therefore, the remaining four degrees (i.e., $S \leq 0.20$) were considered for further inspection with respect to relative deviation and relative bumpiness. Final weighted penalized standardized sum of squares were then computed for the five considered weights.

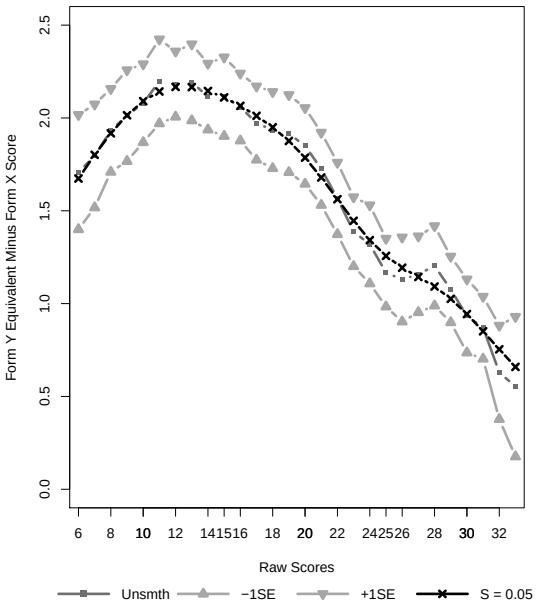
Figure 4 provides difference plots for $S = 0.01$, $S = 0.05$, $S = 0.10$, and $S = 0.20$. Based on Figure 4(a), it can be observed that $S = 0.01$ gave equated-equivalents

Figure 4: Operational Dataset: Raw-to-raw equivalents for postsmoothing, $S$ = 0.01,0.05,0.10, and 0.20

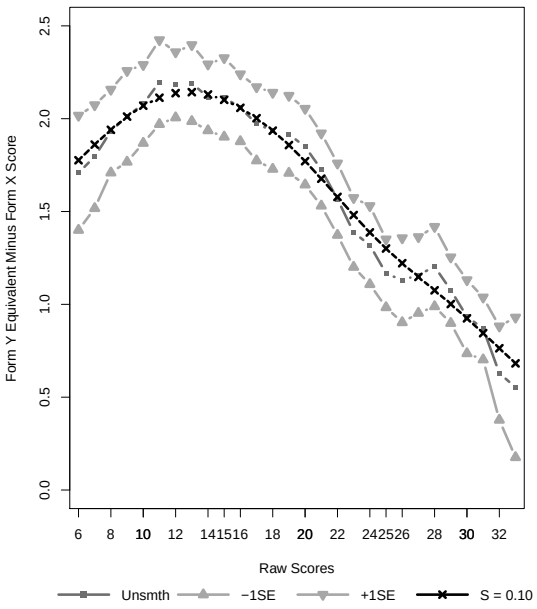
(a) S = 0.01




(b) S = 0.05




(c) S = 0.10




(d) S = 0.20

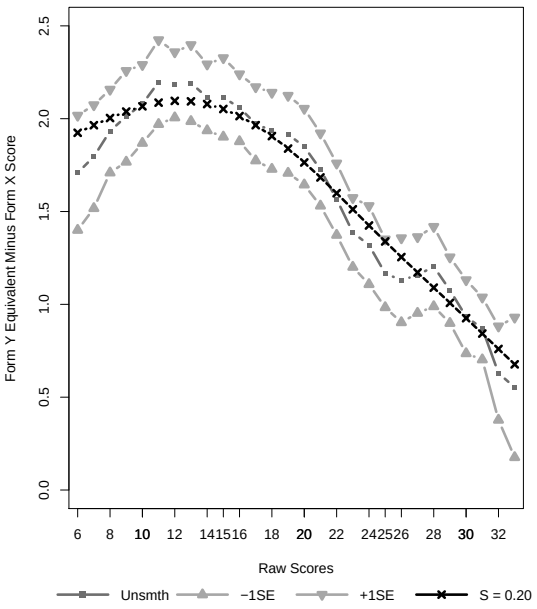


Table 3
*Operational Dataset: Weighted Penalized Standardized Sum of Squares for Various Pairs of Weights*

| $(w_D, w_B)$ | S=0.01 | S=0.05 | S=0.10 | S=0.20 | S=0.30 | S=0.50 | S=0.75 | S=1.00 |
|---|---|---|---|---|---|---|---|---|
| (1, 0) | **0.0151** | 0.0531 | 0.0963 | 0.1778 | NA | NA | NA | NA |
| (2/3, 1/3) | 0.1441 | **0.1022** | 0.1044 | 0.1433 | NA | NA | NA | NA |
| (1/2, 1/2) | 0.2087 | 0.1267 | **0.1084** | 0.1260 | NA | NA | NA | NA |
| (1/3, 2/3) | 0.2733 | 0.1512 | 0.1124 | **0.1087** | NA | NA | NA | NA |
| (0, 1) | 0.4024 | 0.2003 | 0.1205 | **0.0742** | NA | NA | NA | NA |

*Note*: NA implies that there is at least one score outside the 1SE band. For each pair of weight, a boldfaced number refers the minimum $WP(S)$ among different $S$s excluding those for NA.

that were quite bumpy and close to the unsmoothed relationship; as $S$ increased, the relationships became smoother and deviated farther from the unsmoothed relationship.

Table 3 shows the weighted penalized standardized sum for various weights and smoothing degrees. Table 3 shows that the optimum $S$ depended on the choices of weights. As mentioned above, degrees $S \geq 0.30$ had at least one score whose equated-equivalent was outside the 1SE band. Thus, for each weighting scheme, degrees from 0.01 to 0.20 were considered as candidates for the optimum degree. According to Table 3, when full weight was given to relative deviation (i.e., $w_D = 1$), $S = 0.01$ was the optimum choice. This result was expected because, when reducing the relative deviation is the main interest, the degree with the least smoothing is expected to give an equating relationship that is the closest to the unsmoothed relationship. For the considered dataset, $S = 0.20$ was the optimum choice as long as $w_B \geq 2/3$. When $w_D = 2/3$ and $w_B = 1/3$, $S = 0.05$ was selected as the optimum degree. When $w_D = w_B = 1/2$, $S = 0.10$ was selected as the optimum degree.

## 4.2 Simulation Study

This section consists of three subsections. The first subsection presents the weighted overall statistics for different weighting schemes. The second subsection presents the effect of test length, and the third section describes the effect of sample size.

### 4.2.1 Overall Results

Table 4 presents the weighted overall statistics for different weights $(w_D, w_B)$, aggregated and averaged over the conditions for test length and sample size. Based on Table 4, the WRMSB (i.e., bias) was smallest when full weight was given to the relative deviation, which is as expected. The WSE (i.e., random error) tended to decrease as $w_B$ increased; it became the smallest for $w_B = 1$. With respect to the WRMSE, $(w_D, w_B) = (0, 1)$ gave the smallest value among all considered weights.

Based on the results, it seems as if weighting schemes did not make noticeably large differences in the WRMSB, WSE, and WRMSE. Note that, for each pair of weights, the weighted overall statistics were obtained for the equating results using the optimum

Table 4
*Weighted Overall Statistics*

| $(w_D, w_B)$ | WRMSB | WSE | WRMSE |
|---|---|---|---|
| (1, 0) | **0.0143** | 0.3477 | 0.3481 |
| (2/3, 1/3) | 0.0167 | 0.3305 | 0.3313 |
| (1/2, 1/2) | 0.0297 | 0.3237 | 0.3250 |
| (1/3, 2/3) | 0.0277 | 0.3188 | 0.3211 |
| (0, 1) | 0.0411 | **0.3157** | **0.3199** |

*Note*: A boldfaced number refers to the minimum $WP(S)$ among different weights $(w_D, w_B)$.

Table 5
*Frequency Distributions of Postsmoothing Degree $S^*$ for Different Weights*

| | $(w_D, w_B)$ | | | | |
|---|---|---|---|---|---|
| | $(1, 0)$ | $(2/3, 1/3)$ | $(1/2, 1/2)$ | $(1/3, 2/3)$ | $(0, 1)$ |
| $S = 0.01$ | 4473 | 32 | 30 | 30 | 30 |
| $S = 0.05$ | 26 | 1610 | 449 | 215 | 183 |
| $S = 0.10$ | 1 | 2077 | 1937 | 1080 | 589 |
| $S = 0.15$ | 0 | 529 | 1083 | 1346 | 829 |
| $S = 0.20$ | 0 | 156 | 484 | 761 | 851 |
| $S = 0.25$ | 0 | 54 | 265 | 453 | 693 |
| $S = 0.30$ | 0 | 19 | 128 | 271 | 513 |
| $S = 0.35$ | 0 | 10 | 58 | 134 | 300 |
| $S = 0.40$ | 0 | 6 | 32 | 84 | 193 |
| $S = 0.45$ | 0 | 2 | 18 | 51 | 107 |
| $S = 0.50$ | 0 | 0 | 7 | 51 | 163 |
| $S = 0.75$ | 0 | 3 | 7 | 21 | 31 |
| $S = 1.00$ | 0 | 2 | 2 | 3 | 18 |
| Total | 4500 | 4500 | 4500 | 4500 | 4500 |

degrees. Using the best choice of all for each scenario might have resulted in similar differences in the weighted overall statistics. Based on Table 4, the WSEs were larger than the WRMSBs, and the WRMSE seems to be affected more by the WSE than by the WRMSB. It may be because the random groups design was considered throughout this study. If the common-item nonequivalent groups design (CINEG) is considered with other study factors such as the proportion of common items, then the direction of magnitudes between the WSE and the WRMSB and the influence to the WRMSE might be the opposite.

Table 5 presents the overall frequency distribution of the optimum degree $S^*$ for different weights. Based on Table 5, for $w_D = 1$ and $w_B = 0$, $S = 0.01$ were selected as the optimum smoothing degree most frequently, as expected. And, as $w_D$ decreased and $w_B$ increased, larger degrees started to be selected as the optimum.

Table 6
*Weighted Overall Statistics for Different Test Lengths Averaged over Sample Size Conditions*

| | WRMSB | | | WSE | | | WRMSE | | |
|---|---|---|---|---|---|---|---|---|---|
| $w_D$ | TL 20 | TL 40 | TL 60 | TL 20 | TL 40 | TL 60 | TL 20 | TL 40 | TL 60 |
| 1 | **0.011** | **0.012** | 0.019 | 0.164 | 0.361 | 0.518 | 0.165 | 0.361 | 0.518 |
| 2/3 | 0.021 | 0.012 | **0.017** | 0.154 | 0.342 | 0.495 | 0.156 | 0.342 | 0.495 |
| 1/2 | 0.029 | 0.013 | 0.019 | 0.151 | 0.335 | 0.485 | **0.155** | 0.335 | 0.486 |
| 1/3 | 0.037 | 0.018 | 0.028 | 0.150 | 0.329 | 0.477 | 0.155 | 0.330 | 0.478 |
| 0 | 0.042 | 0.036 | 0.046 | **0.150** | **0.327** | **0.471** | 0.156 | **0.329** | **0.474** |

*Note*: For each test length, a boldfaced number refers the minimum $WP(S)$ among different weights $w_D$ for which $w_B = 1 - w_D$. TL = Test Length

### 4.2.2 Effect of Test Length

Table 6 presents the weighted overall statistics for the three different test lengths. With respect to the WRMSB, $w_D = 1$ had the smallest value for test length of 20 items, and $w_D = 2/3$ gave the smallest values for test lengths of 40 and 60 items. However, the differences in the WRMSBs between $(0, 1)$ and $(1/3, 2/3)$ for $(w_D, w_B)$ were not very large. For the WSE, the results were the same as those found based on the overall results; i.e., $w_B = 1$ gave the smallest WSEs. For the WRMSE, considering the relative deviation and the relative bumpiness with equal weights gave the smallest WRMSE for test length of 20 items, whereas $(0, 1)$ for $(w_D, w_B)$ resulted in the smallest WRMSE for test lengths of 40 and 60 items. Similar to the overall results, Table 6 shows that the differences in the weighted overall statistics were not large across different weights.

Table 7 presents the frequency distributions of the optimum degree for test lengths of 20, 40, and 60 items for various weighting schemes. For the test length of 20 items, Table 7 shows that small degrees such as 0.01 tended to be optimal if reducing relative deviation was the main interest (i.e., $w_D = 1$ and $w_B = 0$). As interest shifts from reducing relative deviation to reducing relative bumpiness (i.e., $w_D \downarrow 0$ and $w_B \uparrow 1$), larger degrees were selected as the optimum more frequently. Similar patterns were found for test lengths of 40 and 60 items.

However, the pattern was less evident as test length increased, suggesting that choices of weighting schemes mattered more for shorter tests. Note that, since a shorter test has a smaller number of score points, it is typical for observed score distributions and equated relationships to be smoother for shorter tests (all other things being equal). This implies that, for shorter tests, larger degrees of $S$ can satisfy the 1SE band criterion; and, as $w_B \uparrow$, shorter tests have high ceiling for a larger $S$ to be selected as optimum.

Therefore, for each weighting scheme, larger (smaller) values tended to be selected less (more) frequently as the optimum degree as test length increased. For example, for the test length of 20 items, when $w_D = 0$ and $w_B = 1$ were considered, most of the optimum degrees occurred for $S = 0.10$ to $S = 0.50$ (considering frequencies larger than 50 only); in addition, $S = 1.00$ was selected as the optimum degree for 16 cases. For

Table 7
*Frequency Distributions of Postsmoothing Degree $S^*$ for Different Weights and Different Test Length Conditions*

| $(w_D, w_B)$ | (1,0) | | | (2/3,1/3) | | | (1/2,1/2) | | | (1/3,2/3) | | | (0,1) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test Length | 20 | 40 | 60 | 20 | 40 | 60 | 20 | 40 | 60 | 20 | 40 | 60 | 20 | 40 | 60 |
| $S = 0.01$ | 1473 | 1500 | 1500 | 18 | 8 | 6 | 16 | 8 | 6 | 16 | 8 | 6 | 16 | 8 | 6 |
| $S = 0.05$ | 26 | 0 | 0 | 148 | 608 | 854 | 36 | 164 | 249 | 17 | 93 | 105 | 16 | 72 | 95 |
| $S = 0.10$ | 1 | 0 | 0 | 649 | 814 | 614 | 208 | 789 | 940 | 66 | 446 | 568 | 51 | 212 | 326 |
| $S = 0.15$ | 0 | 0 | 0 | 444 | 59 | 26 | 383 | 424 | 276 | 206 | 565 | 575 | 123 | 341 | 365 |
| $S = 0.20$ | 0 | 0 | 0 | 150 | 6 | 0 | 374 | 83 | 27 | 313 | 248 | 200 | 213 | 316 | 322 |
| $S = 0.25$ | 0 | 0 | 0 | 50 | 4 | 0 | 243 | 20 | 2 | 322 | 91 | 40 | 270 | 228 | 195 |
| $S = 0.30$ | 0 | 0 | 0 | 18 | 1 | 0 | 121 | 7 | 0 | 241 | 26 | 4 | 279 | 132 | 102 |
| $S = 0.35$ | 0 | 0 | 0 | 10 | 0 | 0 | 58 | 0 | 0 | 126 | 8 | 0 | 171 | 77 | 52 |
| $S = 0.40$ | 0 | 0 | 0 | 6 | 0 | 0 | 28 | 4 | 0 | 76 | 7 | 1 | 135 | 41 | 17 |
| $S = 0.45$ | 0 | 0 | 0 | 2 | 0 | 0 | 17 | 1 | 0 | 46 | 4 | 1 | 73 | 25 | 9 |
| $S = 0.50$ | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 48 | 3 | 0 | 112 | 41 | 10 |
| $S = 0.75$ | 0 | 0 | 0 | 3 | 0 | 0 | 7 | 0 | 0 | 20 | 1 | 0 | 25 | 5 | 1 |
| $S = 1.00$ | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 16 | 2 | 0 |
| Total | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |

test lengths of 40 and 60 items, the optimum degrees tended to occur for values from $S = 0.05$ to 0.35 with frequencies larger than 50. for a test length of 60 items, smaller degrees tended to be selected as the optimum more frequently than for a test length of 40 items. Similar patterns were found for the other weighting schemes.

### 4.2.3 Effect of Sample Size

Table 8 presents weighted overall statistics for different sample sizes. With respect to WRMSB, $w_D = 1$ gave the smallest values, whereas $w_D = 0$ gave the smallest WSEs. With respect to WRMSE, $w_D = 0$ gave the smallest value for sample sizes of 500 and 3000, whereas $w_D = 1/3$ gave the smallest values for a sample size of 6000. However, as shown in Table 8 the differences in the weighted overall statistics for different weighting schemes were not large.

Table 9 presents the frequency distributions of $S$ for sample sizes of 500, 3000, and 6000 for various weighting schemes. Based on Table 9, for each sample size, as $w_D$ decreased and $w_B$ increased, larger degrees were selected as optimum more frequently. For example, for the sample size of 6000, the smoothing degree of 0.01 tended to be optimal when the main interest was reducing relative deviation (i.e., $w_D = 1$). Larger degrees were selected as optimum more frequently as $w_D \downarrow$; e.g., $S = 0.10$, 0.10, 0.15, and 0.20 were selected as optimum most frequently for $w_D = 2/3$, 1/2, 1/3, and 0, respectively.

Table 9 also shows that smaller degrees tended to be selected as optimum less frequently as sample size increased. For example, for $w_D = 0$ and $w_B = 1$, the number of cases where small degrees such as $S \leq 0.15$ were selected as optimum decreased as

Table 8
*Weighted Overall Statistics for Different Sample Sizes Averaged Over Test Length Conditions*

| | WRMSB | | | WSE | | | WRMSE | | |
|---|---|---|---|---|---|---|---|---|---|
| $w_D$ | 500 | 3000 | 6000 | 500 | 3000 | 6000 | 500 | 3000 | 6000 |
| 1 | **0.028** | **0.014** | **0.009** | 0.617 | 0.251 | 0.176 | 0.617 | 0.251 | 0.176 |
| 2/3 | 0.021 | 0.016 | 0.012 | 0.586 | 0.238 | 0.167 | 0.587 | 0.239 | 0.168 |
| 1/2 | 0.025 | 0.019 | 0.018 | 0.574 | 0.233 | 0.164 | 0.575 | 0.334 | 0.165 |
| 1/3 | 0.032 | 0.027 | 0.025 | 0.566 | 0.229 | 0.161 | 0.567 | 0.234 | **0.164** |
| 0 | 0.043 | 0.042 | 0.039 | **0.560** | **0.228** | **0.160** | **0.562** | **0.232** | 0.165 |

*Note*: For each test length, a boldfaced number refers the minimum $WP(S)$ among different weights $w_D$ for which $w_B = 1 - w_D$. SS = Sample Size

sample size increased. Note that, for a fixed test length, observed score distributions and equating relationships become smoother for larger sample sizes. And, for a smoother equating relationship, smaller degrees tended to be selected as optimum less frequently.

Based on Table 9, larger degrees were selected as optimum less frequently as sample size increased. For $w_D = 0$ and $w_B = 1$, the number of cases where large degrees such as $S \geq 0.50$ were selected as optimum decreased as sample size increased. Note that the standard error bands are smaller for larger sample sizes, implying that the largest degree satisfying the 1SE band criterion becomes smaller as sample size increases. This explains why larger degrees were selected as optimum less frequently for larger sample sizes. Similar patterns were observed for the other weighting schemes – different ranges of smoothing degrees with such patterns were observed for different weights.

# 5 Summary and Discussion

It is desirable for the choice of smoothing degree to result in a smooth function without departing too much from the unsmoothed equivalents. In choosing the optimum degree, visual inspection has been a conventional tool, but is prone to subjectivity. Unfortunately, when there are many equatings to be done and optimum degrees to be selected for the cubic-spline postsmoothing method in a short period of time, which can happen in an operational setting, it is time-consuming and almost impossible to visually inspect all of the difference plots.

Moreover, the visual inspection cannot guarantee that the amount of attention given to deviation and smoothness are consistent across multiple equatings. Based on visual inspection, some might argue that, for example, $S = 0.10$ gave equated-equivalents that did not deviate much from the unsmoothed equated-equivalents and that were smooth enough. Also, some might suggest that additional smoothing would not harm the final results. It is not quite clear how each choice can be justified in terms of how much attention should be given to the deviation and smoothness. The issue will continue to occur as more equatings are performed, so it would be hard to argue that

Table 9
*Frequency Distributions of Postsmoothing Degree $S^*$ for Different Weights and Sample Size Conditions*

| $(w_D, w_B)$ | (1, 0) | | | (2/3, 1/3) | | | (1/2, 1/2) | | | (1/3, 2/3) | | | (0, 1) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample Size | 500 | 3000 | 6000 | 500 | 3000 | 6000 | 500 | 3000 | 6000 | 500 | 3000 | 6000 | 500 | 3000 | 6000 |
| $S = 0.01$ | 1499 | 1496 | 1478 | 20 | 6 | 6 | 19 | 6 | 5 | 19 | 6 | 5 | 19 | 6 | 5 |
| $S = 0.05$ | 1 | 4 | 21 | 444 | 560 | 606 | 143 | 152 | 154 | 100 | 63 | 52 | 98 | 51 | 34 |
| $S = 0.10$ | 0 | 0 | 1 | 758 | 683 | 636 | 593 | 660 | 684 | 351 | 353 | 376 | 279 | 159 | 151 |
| $S = 0.15$ | 0 | 0 | 0 | 181 | 167 | 181 | 399 | 377 | 307 | 414 | 470 | 462 | 291 | 275 | 263 |
| $S = 0.20$ | 0 | 0 | 0 | 51 | 55 | 50 | 166 | 144 | 174 | 264 | 259 | 238 | 264 | 301 | 286 |
| $S = 0.25$ | 0 | 0 | 0 | 24 | 15 | 15 | 76 | 84 | 105 | 139 | 148 | 166 | 169 | 250 | 274 |
| $S = 0.30$ | 0 | 0 | 0 | 6 | 9 | 4 | 43 | 43 | 42 | 72 | 96 | 103 | 126 | 188 | 199 |
| $S = 0.35$ | 0 | 0 | 0 | 6 | 2 | 2 | 28 | 14 | 16 | 49 | 42 | 43 | 82 | 96 | 122 |
| $S = 0.40$ | 0 | 0 | 0 | 5 | 1 | 0 | 13 | 11 | 8 | 29 | 28 | 27 | 54 | 64 | 75 |
| $S = 0.45$ | 0 | 0 | 0 | 1 | 1 | 0 | 6 | 7 | 5 | 18 | 16 | 17 | 24 | 43 | 40 |
| $S = 0.50$ | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 1 | 0 | 24 | 16 | 11 | 57 | 56 | 50 |
| $S = 0.75$ | 0 | 0 | 0 | 3 | 0 | 0 | 7 | 0 | 0 | 19 | 2 | 0 | 25 | 6 | 0 |
| $S = 1.00$ | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 2 | 1 | 0 | 12 | 5 | 1 |
| Total | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 | 1500 |

the same amount of attention is given to the deviation and smoothness throughout the whole equatings. Therefore, there has been a need for the development of a statistical criterion to choose an appropriate smoothing degree for operational use in cubic-spline postsmoothing.

In order to create a statistical criterion, this study considered three conditions: (1) 1SE band, (2) relative deviation, and (3) relative bumpiness, as suggested by Kolen and Brennan (2014). The study has demonstrated that the new statistical criterion for assessing the fitness of cubic-spline postsmoothing functioned appropriately for the two real datasets.

The results of the simulation study, for the most part, met the expectations listed in Section 3.2.2. As test length increased, larger (smaller) values were selected less (more) frequently as the optimum smoothing degree. As sample size increased for a fixed test length, smaller values were selected less frequently as the optimum smoothing degree. As sample size increased, larger values were also selected less frequently as the optimum smoothing degree because of smaller standard errors. The previous two statements may sound contradictory to each other; however, it is reasonable considering that, for a fixed length, observed-score distributions and equating relationship become smoother for larger sample sizes and that standard errors bands are smaller for larger sample sizes.

Based on the overall statistics, the WRMSB tended to become smaller as more attention was given to relative deviation. As the attention shifted from reducing the relative deviation to reducing the relative bumpiness, the WSE tended to become smaller as expected. In addition, the frequency tables showed that, for $w_D = 1$ and $w_B = 0$, $S = 0.01$ tended to be selected as the optimum smoothing degree most frequently. As $w_D \downarrow 0$ and $w_B \uparrow 1$, larger degrees were selected as the optimum.

Based on the simulation study, the weighted overall statistics did not differ much for different weighting schemes for $(w_D, w_B)$. This may be mainly due to the fact that equating results using the optimum smoothing degrees were always used to obtain the statistics. However, weighting scheme mattered in terms of how frequently each degree was selected as the optimum, especially for shorter test lengths (i.e., 20 items). For example, for test length of 20 items, if sole interest is to reduce the relative bumpiness only, the final optimum degrees (i.e., the largest $S$ satisfying the 1SE band condition) can be larger than degrees for which sufficient smoothness is already achieved, which consequently can result in large deviations from unsmoothed equated-equivalents. Therefore, it is important to determine how much attention should be given to the relative deviation and relative smoothness, especially for short tests.

The statistical criterion proposed here leaves open the question about which weighting scheme should be considered. Theoretically, weights for relative deviation and relative bumpiness can be any non-negative numbers as long as their sum is equal to one. However, the authors suggest avoiding giving full weight to either relative deviation or relative bumpiness. When postsmoothing is applied to an equating relationship, it means that the equating relationship after smoothing should be smooth in some sense; and, giving full weight to relative deviation only is contradictory to the concept, implying $w_B \neq 0$. Similarly, postsmoothing should result in equating relationships that do not deviate too much from unsmoothed relationships; and, giving full weight to relative bumpiness should be avoided, implying $w_D \neq 0$. Therefore, the authors suggest that weights for both relative deviation and relative bumpiness be positive in practice.

Throughout the study conditions, the WSE was larger than the WRMSB, and the WRMSE seemed to be affected more by the WSE than by the WRMSB. It may be due to the use of the random groups design in this study. Therefore, future research can consider using the CINEG design with other study factors such as the proportion of common items and group differences in their ability levels. If the CINEG design is considered, the direction of magnitudes between the WSE and the WRMSB and the influence to the WRMSE might not be the same as observed in this study. This current study considers the 1SE band criterion to exclude extremely large values from possible candidates for an optimum degree. Future research can consider loosening the 1SE band criterion so that an optimum degree can be selected based on the deviation and smoothness criteria only.

The statistical criterion proposed in this study is certainly an improvement because the same weighting scheme for deviation and smoothness can be used across multiple equatings, whereas visual inspection cannot guarantee that consistent amount of attention is given to deviation and smoothness across multiple equatings. Therefore, the authors suggest that the statistical criterion presented in this paper be considered for use in postsmoothing.

## References

ACT. (2014). *ACT technical manual*. Iowa City, IA: Author

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*, 3-14.

College Board. (2016). *SAT technical manual: Characteristics of the SAT*. New York, NY: Author.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics*, *30*, 589-600.

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

# APPENDIX

According to Kolen and Brennan (2014), an optimum postsmoothing degree $S^*$ should result in (a) smoothed equated-equivalents that (b) do not deviate too much from unsmoothed equated-equivalents, with the practical constraint that (c) the smoothed relationship should lie mainly within the one standard error (SE) band. The three criteria are associated with equated-equivalents rather than differences between Form Y equivalents and Form X scores (i.e., $e_Y(x_i) - x_i$, difference-scores) that are typically used in determining $S^*$. The use of difference-scores here is motivated primarily by the fact that they much more dramatically reveal bumpiness than do the equated-equivalents themselves. As discussed next, however, difference scores introduce an algebraic complexity.

Suppose that equated-equivalents were used to obtain relative bumpiness. Then, the numerator of Equation (2) becomes:

$$|\{\hat{e}_{Y,S}(x_{i+1}) - \hat{e}_{Y,S}(x_i)\} - \{\hat{e}_{Y,S}(x_i) - \hat{e}_{Y,S}(x_{i-1})\}|; \tag{10}$$

whereas, when difference-scores in Equation (2) are expressed in terms of equated-equivalents, the numerator of Equation (2) becomes:

$$\begin{aligned}
&|\{\hat{dc}_{Y,S}(x_{i+1}) - \hat{dc}_{Y,S}(x_i)\} - \{\hat{dc}_{Y,S}(x_i) - \hat{dc}_{Y,S}(x_{i-1})\}| \\
&= |[\{\hat{e}_{Y,S}(x_{i+1}) - (x_i + 1)\} - \{\hat{e}_{Y,S}(x_i) - x_i\}] - [\{\hat{e}_{Y,S}(x_i) - x_i\} - \{\hat{e}_{Y,S}(x_{i-1}) - (x_i - 1)\}]| \\
&= |\{\hat{e}_{Y,S}(x_{i+1}) - \hat{e}_{Y,S}(x_i) - 1\} - \{\hat{e}_{Y,S}(x_i) - \hat{e}_{Y,S}(x_{i-1}) - 1\}|.
\end{aligned} \tag{11}$$

Compared to Equation (10), Equation (11) includes extra $-1$ terms inside of the absolute-value bars. The same issue occurs in the denominator. Consequently, relative bumpiness computed using difference-scores is algebraically different from relative bumpiness computed using equated-equivalents. Importantly, however, the order of smoothing degrees is invariant whether equated-equivalents or difference-scores are used for computing relative bumpiness.

When a number of equating functions are drawn and compared within one plot, it is difficult to ascertain differences between the functions (Kolen & Brennan, 2014). Each equating relationship is a monotonically non-decreasing function, which makes it more difficult to compare equating functions with respect to bumpiness. Thus, Kolen and Brennan (2014) uses differences between Form Y equivalents and Form X scores to draw

difference-type plots. Such plots facilitate comparing equating functions and choosing optimum smoothing degrees.

In this study, the new statistic is developed so that visual inspection of difference-type plots can be performed analytically. Therefore, despite of the confusion created by using difference-scores, the authors argue that this newly suggested statistic for postsmoothing is congruent with how deviation and smoothness of equating functions are examined in Kolen and Brennan (2014).