Article

Applied Psychological Measurement
2020, Vol. 44(4) 296–310

DOI: 10.1177/0146621619886050
journals.sagepub.com/home/apm
SAGE

Evaluating Robust Scale Transformation Methods With Multiple Outlying Common Items Under IRT True Score Equating

Yong He[1] and Zhongmin Cui[1]

Abstract

Item parameter estimates of a common item on a new test form may change abnormally due to reasons such as item overexposure or change of curriculum. A common item, whose change does not fit the pattern implied by the normally behaved common items, is defined as an outlier. Although improving equating accuracy, detecting and eliminating of outliers may cause a content imbalance among common items. Robust scale transformation methods have recently been proposed to solve this problem when only one outlier is present in the data, although it is not uncommon to see multiple outliers in practice. In this simulation study, the authors examined the robust scale transformation methods under conditions where there were multiple outlying common items. Results indicated that the robust scale transformation methods could reduce the influences of multiple outliers on scale transformation and equating. The robust methods performed similarly to a traditional outlier detection and elimination method in terms of reducing the influence of outliers while keeping adequate content balance.

Keywords

equating, item response theory, multiple outliers, robust scale transformation

Introduction

In many large-scale testing programs, alternate forms are typically used to ensure security and integrity of a test. The small difference in difficulty between two test forms is adjusted through equating to make sure the scores obtained from both test forms are comparable. When there are common items between the two test forms, the item response theory (IRT) can be employed for equating. The common item set, which should represent the content and statistical specifications of the entire set of items, plays a vital role in placing item parameter estimates from the two test forms on the same scale via a scale transformation method (Kolen & Brennan, 2014). For the scale transformation method to work, it is assumed that the common items perform similarly on both test forms and the parameter estimates only differ because of indeterminacy of the IRT scale and sampling error. If the parameter estimates differ because of other reasons,

[1]ACT, Inc., Iowa City, IA, USA

Corresponding Author:
Yong He, ACT, Inc., 500 ACT Drive, Iowa City, IA 52243-0168, USA.
Email: Yong.He@act.org

such as curriculum change or item overexposure, including the outlying common items (or outliers in short) may distort scale transformation and thus undermine the accuracy of equating results (Cook & Eignor, 1991). Therefore, the outliers should be treated before the scale transformation and equating procedures are conducted.

The research literature has shown many efforts in solving the problem with detecting and eliminating of outlying common items before the scale transformation and equating (e.g., DeMars, 2004; Donoghue & Isham, 1998; Guo, Zheng, & Chang, 2015; Holland & Thayer, 1988; Huynh & Meyer, 2010; Raju, 1990; Veerkamp & Glas, 2000). Results from these studies have shown that the detection and elimination of outlying common items surely improved the stability of scale transformation and increased the accuracy of IRT equating. None of these studies, however, have paid attention to the requirement of content representativeness for common items (i.e., the common item set should mimic the total test in terms of content). Although some researchers (e.g., Gao, Hanson, & Harris, 1999; Hanick & Huang, 2002; Hu, Rogers, & Vukmirovic, 2008; Wolkowitz & Davis-Becker, 2015) found that the content representativeness might not significantly affect equating results, other researchers (e.g., Cook & Petersen, 1987; Holland & Dorans, 2006; Klein & Jarjoura, 1985; Kolen & Brennan, 2014) have found that content representativeness of the common item set had great impact on equating results particularly when the examinee groups were different in ability. For this reason, Hanson and Feinstein (1997) suggested removing outliers only when the practice would not harm the content representativeness. In addition, ignoring content representativeness of the common items (aka anchor items) might violate Standard 5.15 by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014, p. 105):

> *In equating studies that employ an anchor test design, the characteristics of the anchor set and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented.*

Methodologically, the outlier elimination methods divide the item set into two exclusive categories: either an outlier or a normally behaved item, depending on whether its item statistics have changed significantly. Whether an item is an outlier or not is typically determined by practitioners, and the procedure is somewhat subjective—practitioners may not agree with each other on the criterion. In other words, it is difficult to obtain a consistent criterion because the outlying behavior of an item is not simply ''yes'' or ''no'' but a matter of degree. In reality, an item can deviate in various degrees from the pattern assumed by the normally behaved items. To take the degree of deviation into account, He, Cui, and Osterlind (2015) proposed two robust scale transformation methods, the *Area Weighted* (*AW*) method and the *Least Absolute Values* (*LAV*) method (see the ''Method'' section for details), by assigning weights to items according to their distance deviated from the assumed pattern. Specifically, small weights are given for outlying items to reduce their influence on scale transformation, but large weights for normally behaved items so that they have a significant contribution to scale transformation. Note that the detection and elimination methods can be treated as a special case of the robust methods where the weight of each item is assigned to be either 1 or 0.

In He et al. (2015), the robust scale transformation methods were shown to be effective in reducing the impact of an outlier on the accuracy of scale transformation and equating while maintaining content representativeness. However, the study was limited to only one outlier in the common item set. In practice, it is not uncommon to see multiple outliers in the common item set. The effect of multiple outliers on scale transformation and equating is much more complicated than the single outlier case. An outlier may not be identified because of the presence of

other adjacent outliers (aka masking effect, Hadi & Simonoff, 1993), or an normally behaved item might be falsely identified as an outlier because some outliers are so influential that they together distort the assumed pattern (aka swamping effect, Ben-Gal, 2005). Both phenomena make accurate detection of an outlier more challenging. Stepwise or sequential methods have been proposed to detect and eliminate multiple outlying items by flagging and excluding one outlier at a time (e.g., Guo et al., 2015; Hadi & Simonoff, 1993). However, these methods may result in severe content imbalance when multiple items are eliminated from the common item set because each content area typically includes only a small number of items. Suppose a 40-item test, which has 10 common items, has three content areas such that each content area in a balanced common item set has three to four items on average. If a given content area has three items and two of them are identified as outliers, the content balance will be broken by eliminating the identified outliers. As a consequence, the authors believe that the widely used outlier detection and elimination methods fall short when facing multiple outliers.

The purpose of this study was to evaluate the robust scale transformation methods when multiple outliers are present in the common item set. A simulation study was conducted based on real item parameter values obtained from a large-scale testing program. IRT true score equating method was used under the common-item nonequivalent groups equating design. The robust scale transformation methods were compared to traditional outlier-handling methods by evaluating the recovery of scale transformation coefficients and equating results. The performance of the robust methods under different test lengths, number of outliers, and strength of outliers was evaluated.

Method

Data

The authors constructed two tests with different lengths using real items from a large-scale achievement test: a long test form with 120 dichotomously scored items including 40 common items and a short test form with 45 dichotomously scored items including 15 common items. Each test included two test forms, old and new. The authors used item parameter values estimated from real data to generate data used in this study. Item responses of 3,000 simulated examinees per test form were generated using R. The authors assumed a normal ability distribution, $\theta \sim N(0, 1)$, for the examinees taking the old form. For the examinees taking the new test forms, they assumed normal ability distribution of $\theta \sim N(0.25, 1.1^2)$ or $\theta \sim N(0.5, 1.2^2)$. The generated item responses were calibrated using the BILOG-MG 3 program (Zimowski, Muraki, Mislevy, & Bock, 2003).

Scale Transformation Methods

Under the three-parameter logistic (3PL) IRT model, the probability for examinee i with ability θ_i to correctly answer item j is

$$p_{ij}(\theta_i; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}},$$

where a_j, b_j, and c_j are the item parameters for item j indicating discrimination, difficulty, and pseudo-guessing, respectively. The constant D is typically set to 1.7 to make the logistic ogive approximate the normal ogive.

To put item parameter estimates of the new form (denoted as F) on the scale of the old form (denoted as T), we need to compute scale transformation coefficients A and B (i.e., slope and intercept of a linear transformation) using the following equations:

$$a_{Tj} = \frac{a_{Fj}}{A}, \qquad b_{Tj} = Ab_{Fj} + B, \qquad c_{Tj} = c_{Fj},$$

where a_{Tj}, b_{Tj}, and c_{Tj} are item parameters on the old form scale, and a_{Fj}, b_{Fj}, and c_{Fj} are item parameters on the new form scale. After putting the ability and item parameter estimates of the new form on the old form scale, we can define the difference between the two probabilities (one on the old form and the other on the new form) for the i^{th} examinee to answer the j^{th} item correctly as

$$d_{ij} = p_{ij}\left(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}\right) - p_{ij}\left(\theta_{Ti}; \tfrac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj}\right),$$

where θ_{Ti} represents ability of examinee i on the old form scale.

With the Stocking-Lord method (Stocking & Lord, 1983), the scale transformation coefficients are obtained through minimizing the following loss function

$$L = \sum_i \left(\sum_j d_{ij} \right)^2.$$

The Stocking-Lord scale transformation was conducted using the computer program ST (Hanson & Zeng, 1995).

By contrast, the robust methods minimize a different loss function

$$L = \sum_i \sum_j w_j d_{ij}^2, \tag{1}$$

where w_j is the weight assigned to item j.

For the *AW* method, the Huber (1981) weights are used to define the weights as

$$w_j = \begin{cases} 1 & |e_j| \le k \\ k/|e_j| & |e_j| > k \end{cases},$$

where k is the tuning constant and was set to 1.345 to obtain an efficiency of 95% when the errors are normally distributed (see Huber for details) and e_j is the standardized area between two item characteristic curves (ICCs) using median absolute deviation (MAD; Wilcox, 2012)

$$e_j = \frac{0.6745 * \sum_q |d_{qj}| * \Delta\theta}{MAD\left(\sum_q |d_{qj}| * \Delta\theta \right)}$$

where q represents the quadrature points of θ between –4 and 4 for item j, d_{qj} is the difference between the two probabilities obtained from the old form and the new form at a quadrature point of θ, the constant of 0.6745 rescales MAD to a normal distribution (see Wilcox for details), and $\Delta\theta$ is the interval of abilities between two adjacent quadrature points.

For the *LAV* method, the weights are defined as $1/|d_{ij}|$. The loss function in equation (1) can then be simplified as

$$L = \sum_i \sum_j |d_{ij}|.$$

Outlier Simulation

The short form included 0, 1, or 3 common items whose parameter values were adjusted, and the long form contained 0, 1, or 3 items whose parameter values were adjusted. In other words, the proportion of outliers in a common item set was 0%, 6.7%, or 20% for the short test form, and 0%, 2.5%, or 7.5% for the long test form. Here, the authors included the conditions of zero or one item with changed *b*-parameter for comparison reasons in this study.

To simulate the magnitudes of item parameter change, the predetermined number of item(s) was (were) randomly selected from the common item set when needed. Both *a*- and *b*-parameters of the selected item(s) were adjusted to mimic practical situations. The change of *a*-parameter values followed a uniform distribution, $\Delta a \sim U(0.1, 0.5)$. The authors simulated two conditions to adjust the *b*-parameter values: (a) *small change* in which the change of *b*-parameter values followed a uniform distribution, $\Delta b \sim U(-0.5, -0.1)$, or (b) *large change* in which the change of *b*-parameter values followed a uniform distribution further away from zero, $\Delta b \sim U(-1.0, -0.5)$. After changing the item parameter values, the authors simulated the item responses for the examinees based on the adjusted item parameter values.

Outlier Treatment Methods

There were five outlier treatment methods used in this study:

1. *No Treatment*. The authors used the entire common item set for scale transformation without any special treatment. This method would show the influence of ignoring outliers on the accuracy of scale transformation, thus providing a baseline for the comparisons.
2. *Elimination*. The *a*- or *b*-parameter estimates of the common items on the new form were compared with those on the old form after being placed on the same scale using the Stocking-Lord method. If the absolute difference of either *a*- or *b*-parameter estimate was larger than 0.5 (He, Cui, Fang, & Chen, 2013), the item was excluded from the common item set. The Stocking-Lord method was rerun after removing all outliers. Note that this step is a one-time screening procedure.
3. *Area Weighted* (*AW*). The aforementioned *Area Weighted* method was applied to all common items to conduct scale transformation.
4. *Least Absolute Values* (*LAV*). The aforementioned *Least Absolute Values* method was applied to all common items to conduct scale transformation.
5. *Raju's Differential Functioning of Items and Tests* (DFIT; Raju, van Der Linden, & Fleer, 1995). The ICCs were compared after the two sets of item parameters had been placed on the same scale using the Stocking-Lord method. The outlier detection procedure was implemented using R package DFIT (Cervantes, 2017), where the Mantel-Haenszel statistic was transformed into the ETS delta scale (Roussos & Stout, 1996). The Stocking-Lord scale transformation was rerun after removing items with large or C-level delta from the common item set.

After dealing with the outliers and placing the item parameter estimates from both new and old test forms on the same scale using one of the five methods, the authors conducted IRT true score equating and evaluated the equating results.

The idea of maintaining content representativeness was one of the major motivations to develop the robust scale transformation methods. At first thought, it might be intuitive to include different conditions of content. However, the authors did not simulate different conditions of content representativeness in this study because (a) the robust methods have the advantage over traditional outlier detection and elimination methods, such that the content representativeness would not be affected during scale transformation and (b) the effects of content representativeness have already been studied. By doing so, the authors limited their focus to examining differences among the outlier treatment methods.

IRT True Score Equating

IRT true score equating was conducted according to Kolen and Brennan (2014) where the number-correct true scores from two test forms are equivalent at a given ability level. In this study, the authors used scores of 1-45 as the true scores for the short test form and 1-120 for the long test form.

Evaluation

The authors replicated the simulation 100 times and computed bias and root mean squared error (RMSE) of the scale transformation coefficients to evaluate the accuracy of different outlier treatment methods.

Let ω denote the true value of a scale transformation coefficient, either A or B, $\hat{\omega}$ represent an estimate of the scale transformation coefficient, $\bar{\omega}$ represent the mean of the scale transformation coefficient over replications, and R represent the number of replications. The bias was computed by

$$Bias(\hat{\omega}) = \bar{\omega} - \omega,$$

and the RMSE was calculated by

$$RMSE(\hat{\omega}) = \sqrt{[SE(\hat{\omega})]^2 + [Bias(\hat{\omega})]^2},$$

where

$$SE(\hat{\omega}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\omega}_r - \bar{\omega})^2}.$$

The bias and RMSE of the equated score at each number-correct score point were also compared to the true equating relationship that was obtained when no outlier was simulated.

Considering all score points, the weighted root mean square error (WRMSE) and weighted absolute bias (WAB) were calculated as follows:

$$WRMSE = \sum_{s=0}^{k} P_s \cdot RMSE_s,$$

$$WAB = \sum_{s=0}^{k} P_s \cdot |Bias_s|,$$

where P_s is the proportion of examinees at score point s and k is the number of the maximum score point on the test.

Table 1. Bias of Scale Transformation Coefficients in the Short Test Form.

		A					*B*				
Outlier condition		NT	*LAV*	*AW*	EL	DFIT	NT	*LAV*	*AW*	EL	DFIT
No outlier		**–0.003**	**–0.003**	**–0.003**	**–0.003**	**–0.003**	–0.001	**0.000**	–0.001	–0.001	–0.001
One	Small	–0.001	–0.004	–0.001	**0.000**	0.004	0.024	**0.007**	0.014	0.021	0.017
	Large	**0.001**	–0.006	0.009	–0.004	0.004	0.053	**0.007**	0.013	0.011	0.009
Three	Small	0.010	**–0.002**	0.016	0.013	0.027	0.069	**0.013**	0.039	0.065	0.047
	Large	–0.011	–0.006	0.040	**–0.004**	**0.004**	0.156	**0.011**	0.036	0.040	0.032

Note. Bold font indicates the best method under the given condition. *A* stands for the slope coefficient of scale transformation, and *B* stands for the intercept coefficient of scale transformation. NT = no treatment; *LAV* = least absolute values; *AW* = area weighted; EL = elimination; DFIT = Raju's Differential Functioning of Items and Tests. Small indicates $\Delta b \sim U(-0.5, 0.1)$, and large indicates $\Delta b \sim U(-1.0, -0.5)$.

For evaluation, the authors anticipated that the robust scale transformation methods would reduce errors in the presence of outlying common items, and their performance would be comparable to other outlier treatment methods, such as the Elimination method and the DFIT method. Inferential statistics have been suggested to extend the descriptive results of the evaluation when effects of multiple factors were investigated and compared in simulation studies (e.g., Donoghue & Isham, 1998; Harwell, 1997; Harwell, Stone, Hsu, & Kirisci, 1996; Yoes, 1995). As a result, the authors conducted analysis of variance (ANOVA) to investigate the relationship between the simulated factors (number of outliers simulated, the magnitude of item parameter change, and outlier treatment methods), for both the short test form and the long test form. Although different ability distributions were simulated for examinees, the authors did not include this condition in the ANOVA procedures because the comparison among the ability distributions was not of the main interest of this study. The log-transformed absolute bias and RMSE were used as the dependent variables for the scale transformation coefficients, and log-transformed WAB and WRMSE were used as the dependent variable for the equated scores because the actual values were positively skewed. The highest order interaction term was omitted due to the fact that there was only one observation within each cell in this study design. Following the ANOVA, the authors interpreted the results to determine statistical significance at $\alpha = .05$ level and practical significance by using omega-squared (ω^2) effect size following the guidelines of Wickens and Keppel (2004). Specifically, $\omega^2 \geq 0.01$ indicates a small effect, $\omega^2 \geq 0.06$ shows a moderate effect, and $\omega^2 \geq 0.15$ means a large effect.

Results

Evaluation of Scale Transformation

To evaluate the recovery of scale transformation coefficients, the bias and RMSE of scale transformation coefficients were computed. Tables 1 and 2 show the results for the short test form.

As can be seen from Table 1, all outlier treatment methods performed equally well regarding bias when there was no outlier in the common item set. The differences among these methods were very small and most likely due to sampling error. When there was one outlier, all outlier treatment methods performed similarly in recovering the slope coefficient (*A)* values, whereas the *LAV* method performed the best in recovering the intercept coefficient (*B*) values. When the change of *b*-parameter was large, the *LAV* method had the least bias for the intercept coefficient as compared to the other methods, although the differences among the outlier treatment methods

Table 2. RMSE of Scale Transformation Coefficients in the Short Test Form.

Outlier condition		A					B				
		NT	*LAV*	*AW*	EL	DFIT	NT	*LAV*	*AW*	EL	DFIT
No outlier		**0.021**	0.023	0.023	**0.021**	**0.021**	**0.026**	**0.026**	**0.026**	**0.026**	**0.026**
One	Small	0.026	**0.022**	0.024	0.025	0.024	0.035	**0.025**	0.029	0.034	0.031
	Large	0.047	**0.023**	0.024	0.027	0.027	0.061	**0.027**	0.030	0.032	0.031
Three	Small	0.039	**0.024**	0.046	0.038	0.042	0.077	**0.027**	0.051	0.074	0.058
	Large	0.077	**0.026**	0.053	0.049	0.053	0.163	**0.026**	0.048	0.062	0.056

Note. Bold font indicates the best method under the given condition. *A* stands for the slope coefficient of scale transformation, and *B* stands for the intercept coefficient of scale transformation. RMSE = root mean squared error; NT = no treatment; *LAV* = least absolute values; *AW* = area weighted; EL = elimination; DFIT = Raju's Differential Functioning of Items and Tests. Small indicates $\Delta b \sim U(-0.5, 0.1)$, and large indicates $\Delta b \sim U(-1.0, -0.5)$.

Table 3. Omega Squares (ω^2) from ANOVA on Bias and RMSE for the Short Test Form.

		Transformation coefficients				Equated scores	
		Bias		RMSE			
Source of variation	DF	*A*	*B*	*A*	*B*	WAB	WRMSE
Outlier treatment methods (*M*)	4	0.048	**0.099**	**0.101**	**0.200**	**0.074**	**0.043**
Number of outliers simulated (*N*)	1	**0.231**	0.002	**0.171**	**0.110**	**0.067**	**0.145**
Magnitude of *b*-parameter changes (*C*)	2	0.046	**0.590**	**0.387**	**0.325**	**0.621**	**0.639**
M×*N*	4	0.044	0.014	0.020	0.031	–0.003	0.000
M×*C*	8	–0.018	0.013	**0.119**	**0.136**	**0.111**	**0.055**
N×*C*	2	**0.149**	**0.241**	**0.119**	**0.109**	**0.100**	**0.079**

Note. Bold font indicates both statistically significant ($p < .05$) and practically significant ($\omega^2 \geq 0.01$). *A* stands for the slope coefficient of scale transformation, and *B* stands for the intercept coefficient of scale transformation. ANOVA = analysis of variance; RMSE = root mean squared error; WAB = weighted absolute bias; WRMSE = weighted root mean square error.

were small. For the condition of three outliers, the *LAV* method performed the best in recovering the transformation coefficients, both *A* and *B*, when the change was small. The *Elimination* method and the DFIT method had the least bias for the slope coefficient *(A)* values when the change was large, which were slightly different from the *LAV* method. Regardless of the magnitude of changes, the *LAV* method yielded the least bias in recovering the intercept coefficient (*B*).

Table 2 shows that all outlier treatment methods yielded similar RMSE values when there was no outlier in the common item set. The differences among these methods were small and most likely due to sampling error. It is consistent with the finding from Table 1 in that all outlier treatment methods performed equally well when there was no outlier in the data. This finding seems to suggest that it is safe to use the robust methods as a general scale transformation method even when no outlier is present in the data. This finding is consistent with what He et al. (2015) reported.

When the magnitude of *b*-parameter change was large, the RMSE yielded by the No Treatment method dramatically increased, while other methods significantly helped to remedy the situation in terms of reducing RMSE of the scale transformation coefficients. Regardless of the number of outliers, the *LAV* method performed the best and yielded the least RMSE values in recovering both slope and intercept coefficients (Table 2).

As expected, the ANOVA results (Table 3) indicated statistically significant differences in bias among the outlier treatment methods. Specifically, the magnitude of *b*-parameter changes

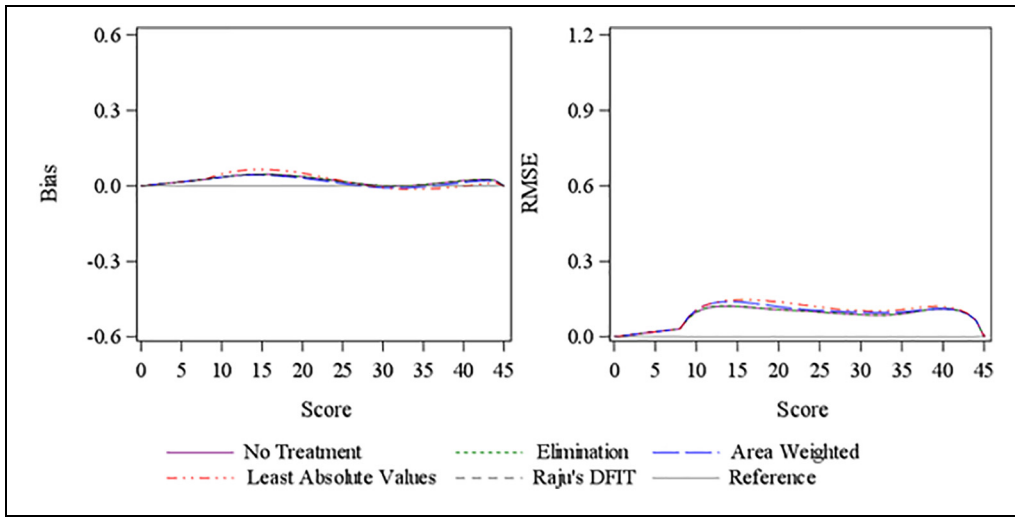

Figure 1. The bias and RMSE of equated scores in the short test form under the *no outlier* condition.
Note. RMSE = root mean squared error; DFIT = differential functioning of items.

(*C*; $\omega^2 = 0.590$) had the largest impact on the recovery of *B* coefficient, and the *LAV* method generally had the least bias among the investigated outlier treatment methods (*M*; $\omega^2 = 0.099$). The ANOVA results also show that differences in RMSE among outlier treatment methods were statistically significant (*M*; $\omega^2 = 0.212$), and the outlier treatment method had significant interaction with the magnitude of *b*-parameter change (*C*; $\omega^2 = 0.440$ for coefficient *A* and $\omega^2 = 0.043$ for coefficient *B*). Full ANOVA tables are given in the online supplement.

Evaluation of Equated Scores

The bias and RMSE of equated scores on the short test form under the *no outlier* condition are shown in Figure 1. This figure shows that all methods yielded similar bias and RMSE. Although the robust methods yielded larger errors as compared to other methods, the difference was small (less than 0.04 for RMSE).

Figure 2 shows the bias of the equated scores on the short test form under different outlier conditions. Consistent with the findings in the literature, the No Treatment method yielded the largest bias in all outlier conditions. The more severe was the outlier(s), the larger was the bias for the equated scores. When the change of *b*-parameter value was small, the robust methods (*AW* and *LAV*) generally yielded smaller bias than the elimination method and the DFIT method. When the change of *b*-parameter value was large, all four outlier treatment methods seemed to work comparably by reducing the bias when there was only one outlier in the data. The Elimination method yielded slightly less bias than the others, but there was not a clear pattern to conclude. For the condition of three outliers changed with a large magnitude, the difference was more evident—the Elimination method, at the most score points, yielded the smallest bias among all the outlier treatment methods. The comparison between the two robust methods was not conclusive because sometimes the *AW* method was better and the other times the *LAV* was better.

Figure 3 shows the RMSE results of the equated scores for the short test form. When the change of *b*-parameter was small, the *LAV* method performed the best by yielding the smallest

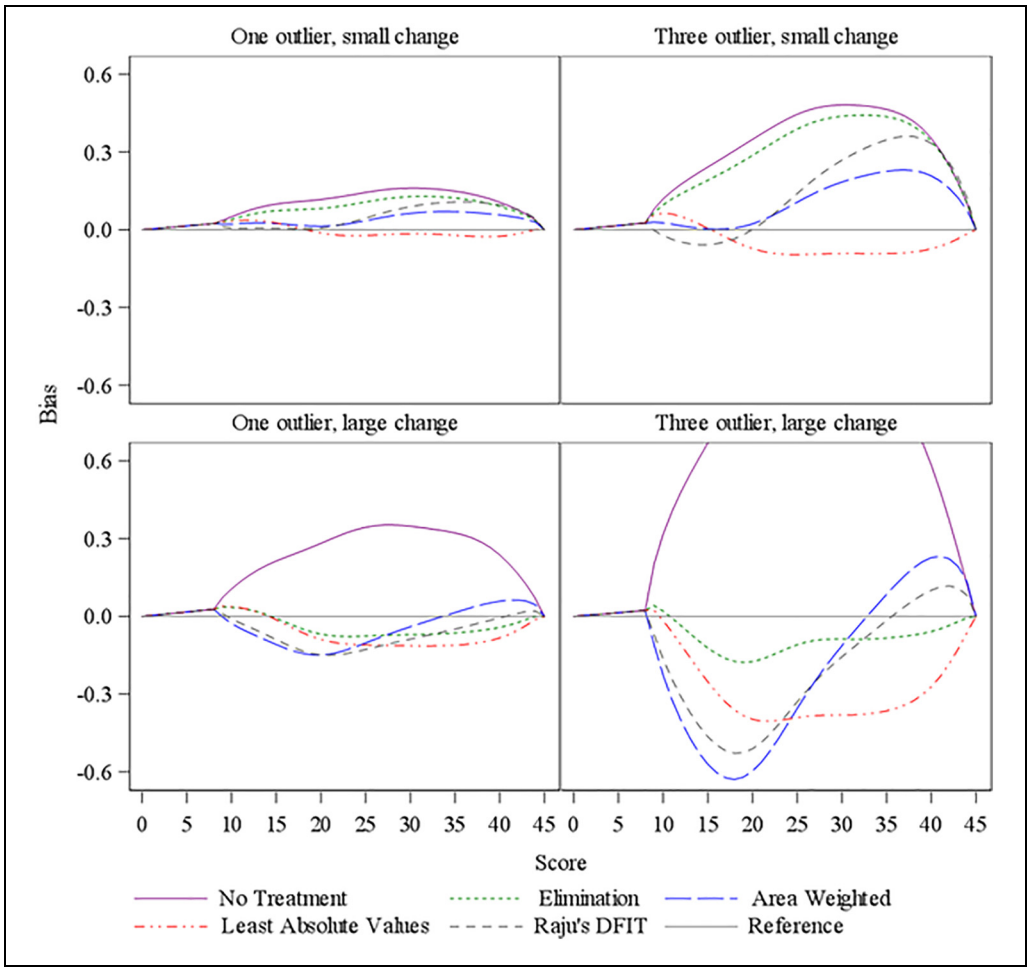

Figure 2. The bias of equated scores in the short test form in the presence of outliers.
Note. DFIT = differential functioning of items.

RMSE values at almost all score points. This finding was more apparent for the condition with multiple outliers. When the change of *b*-parameter was large, the *LAV* method also performed well. It was slightly better than the DFIT method and the Elimination method by yielding smaller RMSE values at almost all score points. The comparison between the two robust methods, similar to the bias results, was inconclusive. Specifically, the *AW* method had smaller RMSE than the *LAV* method in the middle of the score range and larger elsewhere. The results are supported by the ANOVA results for WAB and WRMSE of the equated scores (Table 3), where the main effect of outlier treatment method and its interactions with the magnitude of *b*-parameter change were found to be statistically significant (see the online supplement for full ANOVA results).

Although the comparison between the Elimination method and the DFIT method was not a focus of this study, the results did show some differences regarding their performance in outlier identification and, consequently, scale transformation and equating accuracies. Generally, as seen in Figure 3, the DFIT method outperformed the Elimination method when the change of

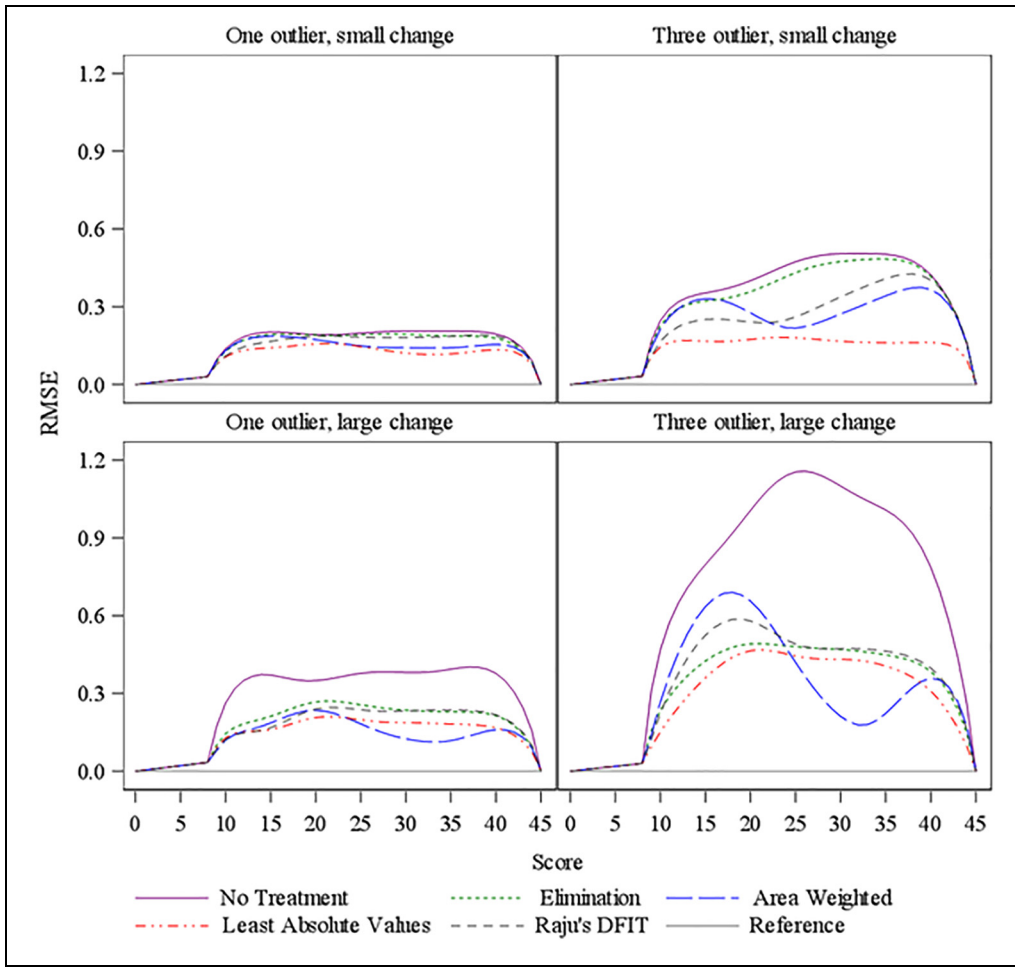

Figure 3. The RMSE of equated scores in the short test form in the presence of outliers.
Note. RMSE = root mean squared error; DFIT = differential functioning of items.

b-parameter was small, and the two methods were similar when the change was large. The finding was also supported by the results of outlier detection rate—the DFIT method had higher outlier detection rate when the change of *b*-parameter was small and similar rate when the change of *b*-parameter was large (results are not shown).

The results for the long test form are provided in the online supplement, showing similar pattern found for the short test form.

Discussion

Outliers in the common item set pose threats to a successful equating because they can potentially distort the equating relationship by reducing the accuracy of scale transformation. Thus, it is important to detect and treat outliers in the common item set before conducting equating. The treatment of outliers is not necessarily equivalent to either elimination or inclusion. He et al. (2015) proposed two robust methods using weights more than zero (as in the case of

elimination) and one (as in the case of inclusion). Their results indicate that the robust methods reduced the influence of outliers on scale transformation and thus on the equating accuracy when only one outlier was in the common item set. The results of He et al. seem promising; however, multiple outliers should be examined before the method can be adapted in practice because (a) multiple outliers are likely to be observed and (b) a method that works for one outlier may not work for multiple outliers because of the masking and swamping effects.

The performance of the robust methods when there was no outlier in the data is comforting because using these methods in place of traditional approaches (e.g., the Stocking-Lord method) seemed to work well. It implies that we do not need to screen the data before applying the robust methods, regardless of the presence/absence of outliers. The Elimination method, the DFIT method, or other outlier detection and elimination methods (He et al., 2013; Raju, 1990) generally have at least three steps: (a) scale transformation, (b) outlier detection and exclusion, and (c) scale transformation with the ''cleaned'' common item set. The stepwise or sequential methods (Guo et al., 2015) have even more steps. Unlike these methods, the robust scale transformation methods are one-step procedures. This feature of the robust methods is appealing in practice because it would be easier and time-saving for operational use compared with the traditional approaches. The performance of the robust methods when there was one outlier was similar to what was reported by He et al. (2015). When there were multiple outliers, the results of the present study confirmed that the same conclusion holds.

In this study, the authors compared the robust methods to the Elimination method and the DFIT method which are widely used in practice but were not studied by He et al. (2015). Although it occasionally yielded larger bias than the Elimination method and the DFIT method, the *LAV* method generally performed better than the two methods by yielding smaller RMSE under almost all conditions. For carefully developed tests, large deviation of *b*-parameters might be handled in the earlier stages. On the other hand, a small change of *b*-parameters might be more commonly seen in practice. The results were again found to favor the *LAV* method due to its capability of reducing errors under such conditions.

Although the *AW* method reduced the influence of outliers compared to the No Treatment method, its performance across all score points appeared to be unstable. For example, the bottom right graph in Figure 3 shows a wavy curve for the *AW* method, which indicates that the *AW* method yielded lower RMSE values than the other methods, mostly at the lower half of the score points but larger RMSE values at the higher half of the score points. For this reason, the authors do not recommend the *AW* method, at least under the conditions researched in this study. However, it does not mean that we should discard the *AW* method. The performance of this method might be affected by the tuning constant. With the right tuning constant, the performance of the *AW* method might be improved.

The overall performance of the *LAV* method is the best among methods investigated in the study, thus the authors recommend it for an operational trial. The *LAV* method has an apparent advantage that it does not need to set up a cutoff point or a constant as required by the other methods. In addition, this method maintains the content balance because it does not require to exclude any item from the common-item set. One might argue that a minimal weight may eventually be equivalent to deleting an item. Strictly speaking, the *LAV* method will not remove an item unless the difference between the two probabilities from the old form and the new form for an examinee to answer a given item correctly ($|d_{ij}|$) is infinite, which is impossible in practice. To show this, the authors simulated an outlier with a large *b*-parameter difference (0.65). With the Elimination method or the DFIT method, the item was excluded for scale transformation. The largest difference, $|d_{ij}|$, was 0.32, and the weight, $1/|d_{ij}| = 3.13$, was far from zero. Although this study's results support using the *LAV* method in practice, one should try it out before operational implementation. Inspection of the bivariate plot of item parameters from the

two calibrations is recommended for close monitoring on its initial use and periodical checking on its regular uses.

Not much research on robust scale transformation has been found in the literature, although it deserves more attention because of its potential to provide more accurate scale transformation results. The present study is limited in scope and could be extended in the following ways. First, one could examine the robust methods using more practical conditions of outliers, such as the one proposed by Han, Wells, and Sireci (2012). Second, although the 3PL IRT model was used in this study, the focus was only on changing *a*- and *b*-parameters. The change of *c*-parameter values may affect the *a*- and *b*-parameter values, and this should be investigated in the future. Other IRT models, for example, two-parameter logistic (2PL) model, graded response model (GRM), or generalized partial credit model (GPCM), could also be considered. Third, one could extend this study by improving the *AW* method. In addition to studying the tuning constant of the Huber weighting function for the *AW* method, one could calculate the area in a restricted range (e.g., $-1 < \theta < 1$), which may be a better indicator regarding the outlying behavior of an item than the area over the whole range. Lastly, the robust methods could be compared with many others (e.g., the Haebara scale transformation method, the concurrent calibration method, or others). With more research on the new methods, we will likely have a better scale transformation which will result in more accurate comparable scores for students.

Author's Note

An earlier version of this study was presented at the 2015 NCME conference.

Acknowledgment

The authors thank Dr. Qing Yi, Dr. J. P. Kim, Dr. John Donoghue, and anonymous reviewers for their constructive comments on this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Yong He https://orcid.org/0000-0003-2014-8208
Zhongmin Cui https://orcid.org/0000-0003-2426-6762

Supplemental Material

Supplemental material for this article is available online.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Ben-Gal, I. (2005). Outlier detection. In O. Maimon & L. Rockach (Eds.), *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers*. Dordrecht, The Netherlands: Kluwer Academic.

Cervantes, V. H. (2017). *DFIT: An R package for the differential functioning of items and tests framework*. Bogota, Colombia: Instituto Colombiano para la Evaluación de la Educación.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, *10*, 37-45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*, 225-244.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*, 265-300.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*, 33-51.

Gao, X., Hanson, B. A., & Harris, D. J. (1999, April). *Effect of using different common item sets under the common item non-equivalent groups design*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Guo, R., Zheng, Y., & Chang, H.-H. (2015). A stepwise test characteristic curve method to detect item parameter drift. *Journal of Educational Measurement*, *52*, 280-300.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, *88*, 1264-1272.

Han, K. T., Wells, C. S., & Sireci, S. G. (2012). The impact of multidirectional item parameter drift on IRT scaling coefficients and proficiency estimates. *Applied Measurement in Education*, *25*, 97-117.

Hanick, P. L., & Huang, C.-Y. (2002, April). *Effects of decreasing the number of common items in equating link item sets*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hanson, B. A., & Feinstein, Z. S. (1997). *Application of a polynomial log linear model to assessing differential item functioning for common items in the common-item equating design* (ACT Research Report Series 97-1). Iowa City, IA: ACT.

Hanson, B. A., & Zeng, L. (1995). *ST: A computer program for IRT scale transformation* [Computer software]. Iowa City, IA: ACT.

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement*, *57*, 266-278.

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101-125.

He, Y., Cui, Z., Fang, Y., & Chen, H. (2013). Using a linear regression method to detect outliers in IRT common item equating. *Applied Psychological Measurement*, *37*, 522-540.

He, Y., Cui, Z., & Osterlind, S. J. (2015). New robust scale transformation methods in the presence of outlying common items. *Applied Psychological Measurement*, *39*, 613-626.

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement*, *32*, 311-333.

Huber, P. J. (1981). *Robust statistics*. New York, NY: Wiley.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, *15*, 1-8.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, *22*, 197-206.

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

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197-207.

Raju, N. S., van Der Linden, W. J., & Fleer, P. F. (1995). An IRT based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement*, *19*, 353-368.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215-230.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, *25*, 373-389.

Wickens, T. D., & Keppel, G. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing (3rd ed.)*. New York, NY: Academic Press.

Wolkowitz, A. A., & Davis-Becker, S. (2015). Evaluating common item block options when faced with practical constraints. *Practical Assessment, Research and Evaluation*, *20*(19). Retrieved from http://pareonline.net/getvn.asp?v=20&n=19

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model* (ASC Technical Report 95-1). Saint Paul, MN: Assessment Systems Corporation.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models* [Computer software]. Chicago, IL: Scientific Software.