New Robust Scale Transformation Methods in the Presence of Outlying Common Items

Applied Psychological Measurement

2015, Vol. 39(8) 613–626

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621615587003

apm.sagepub.com

Yong He¹, Zhongmin Cui¹, and Steven J. Osterlind²

Abstract

Common items play an important role in item response theory (IRT) true score equating under the common-item nonequivalent groups design. Biased item parameter estimates due to common item outliers can lead to large errors in equated scores. Current methods used to screen for common item outliers mainly focus on the detection and elimination of those items, which may lead to inadequate content representation for the common items. To reduce the impact of inconsistency in item parameter estimates while maintaining content representativeness, the authors propose two robust scale transformation methods based on two weighting methods: the Area-Weighted method and the Least Absolute Values (LAV) method. Results from two simulation studies indicate that these robust scale transformation methods performed as well as the Stocking-Lord method in the absence of common item outliers and, more importantly, outperformed the Stocking-Lord method when a single outlying common item was simulated.

Keywords

scale transformation, equating, outlier, robust regression, common item

For security reasons, different test forms are administered on different test dates in large-scale testing programs. Because different test forms usually vary slightly in their statistical characteristics, test developers typically implement test equating procedures to “adjust scores on test forms so that scores on the forms can be used interchangeably” (Kolen & Brennan, 2004, p. 2). The common-item nonequivalent group design, in which two different test forms are embedded with a set of common items, is commonly used in test equating. In the context of item response theory (IRT) equating (Lord, 1982), separately calibrated item parameter estimates for both the new and old forms need to be placed on the same scale via a scale transformation procedure. Due to the indeterminacy property of scale location and spread in IRT models, two sets of separately calibrated item parameters often differ by a linear transformation (Baker & Kim, 2004; De Ayala, 2009). Thus, it is necessary to obtain precise estimates of the intercept and slope for this linear transformation so as to accurately place the two sets of IRT parameters on the same

¹ACT, Inc., Iowa City, IA, USA

²University of Missouri, Columbia, USA

Corresponding Author:

Yong He, ACT, Inc., 500 ACT Drive, Iowa City, IA 52243-0168, USA.

Email: Yong.He@act.org

scale. Marco (1977) described a procedure to estimate the linear relationship using the means and standard deviations of the b -parameter estimates for the common items. By contrast, Loyd and Hoover (1980) proposed using the means of a - and b -parameter estimates for the common items to estimate the linear relationship. Both methods are commonly referred to as moment methods. One potential problem with the moment methods is that they are sensitive to inconsistent item parameter estimates for the common items on the two test forms, and this can lead to a distorted scale transformation and inaccurate equating results (Baker & Al-Karni, 1991; Cook & Eignor, 1991).

Attempts to solve the problem have led researchers to consider two main strategies. Some researchers have used concepts from robust regression to reduce the influence of common item outliers on scale transformations when using moment methods (Bejar & Wingersky, 1981; Linn, Levine, Hastings, & Wardrop, 1981; Stocking & Lord, 1983). Others like Haebara (1980) and Stocking and Lord (1983) have developed characteristic curve methods as alternatives to the robust moment methods. This article will present the application of robust regression ideas to the characteristic curve methods. Weight functions have been used with moment methods, such that smaller weights have been assigned to outlying common items. Linn et al. (1981) defined weights as inversely proportional to the estimated standard error of the estimated item difficulties. Bejar and Wingersky (1981) weighted an item according to the deviance from the transformation line without considering standard errors. Stocking and Lord (1983) proposed an iterative weighted Mean/Sigma method by combining the aforementioned methods. Under the framework of characteristic curve methods, however, all item parameters are simultaneously considered to obtain the intercept and slope of the linear transformation by minimizing a specific loss function. The loss function is defined as the sum of squared differences between the item characteristic curves (ICCs) over examinees in the Haebara method and the sum of squared differences between the test characteristic curves (TCCs) over examinees in the Stocking-Lord method. Ogasawara (2001) found that the characteristic curve methods were more robust against common item outliers than the moment approaches. Similarly, Stocking and Lord (1983) found that the characteristic curve methods usually outperformed the robust scale transformation method based on moments. As a consequence, the characteristic curve methods are predominately used in practice rather than the moment methods.

However, Haebara (1980, p. 149) suggested that “a possible modification of the method to make it more robust to the existence of outliers may be to remove those items . . . from the equating process.” Researchers have proposed several methods to detect and remove outliers in practice: (a) visual inspection of IRT parameter plots (Kolen & Brennan, 2004), (b) the displacement method, and (c) the residual analysis method based on ordinary least squares (OLS) regression (He, Cui, Fang, & Chen, 2013).

These outlier-removing methods, to some extent, improve the stability of the scale transformation and increase the accuracy of IRT equating in the presence of outliers. Removing outliers, however, concerns test developers in that it may weaken content representativeness of the common item set. To represent the group differences, the common item set needs to be a “mini version” of the test forms to be equated. The common item set should “behave similarly in the old and new forms” and “be proportionally representative of the total test forms in content and statistical characteristics” (Kolen & Brennan, 2004, p. 19). Research has shown that content-representative common items have less equating bias than unrepresentative common items (Cook & Petersen, 1987; Klein & Jarjoura, 1985). In other words, eliminating an outlying common item is a sword with two edges: (a) It improves equating accuracy by removing the effect that an outlying item would have on the item parameter estimates, but (b) it distorts equating by undermining content representativeness. Therefore, there is a need for new methods to reduce the impact of inconsistent item parameter estimates while maintaining content representativeness.

The objective of this study was to propose two new scale transformation methods, referred to as the Robust Scale Transformation methods, to reduce the impact of inconsistent item parameter estimates while maintaining content representativeness. The proposed methods combine concepts from robust regression with characteristic curve methods. The authors used simulations to explore two robust scale transformation methods, and they compared the two proposed methods with the Stocking-Lord method. The bias and the root mean square errors (RMSE) of equated scores were calculated to evaluate the effectiveness of the new methods.

Method

In this section, the authors first review the basic concepts of robust regression and then describe how they incorporated the concept of robust regression into the newly proposed scale transformation methods.

Robust Regression

In statistics, the OLS regression method is commonly used to estimate a linear relationship between two variables. The slope and intercept parameter estimates of the linear relationship can be obtained by minimizing the sum of squared deviation scores or residuals (i.e., the difference between observed values and predicted values) given by

$$\sum_i r_i^2,$$

where r_i is the residual of the i th observation. When all assumptions, such as linearity, normality, and homoscedasticity, are met, the OLS regression estimates are consistent. When outliers are present, however, estimates from OLS regression are questionable because of inflated residuals due to outliers. An outlier tends to pull the regression line toward itself and away from the true relationship. In response to this problem, Huber (1973) proposed a robust alternative to OLS regression to deal with the presence of outliers. In his method, a weight function was used to reduce the influence of outliers. Later on, more robust regression methods have been proposed, all of which assign less weight to outlying observations to reduce their influence on the estimation of the regression parameters. In these methods, the slope and intercept parameter estimates of the linear relationship can be obtained by minimizing the sum of weighted squared deviation scores given by

$$\sum_i w_i * r_i^2,$$

where w_i is a weight for the i th observation. There are different ways to define the weight function. To avoid totally eliminating items from the common item set, which is equivalent to assigning 0 weight to an item, the Huber function and the least absolute value (LAV) were used as the weight functions in this study because their weights cannot be 0 unless the deviation is infinite.

Robust Scale Transformation

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

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

where a_j , b_j , and c_j are the item parameters for item j indicating discrimination, difficulty, and pseudo-guessing, and D equals 1.7 so as to make the logistic ogive approximate the normal ogive.

Consider two test forms, with the first being the *to* scale denoted by T (old test form) and the second being the *from* scale denoted by F (new test form). The equations that transform the new test to the old test scale have coefficients A and B such that

$$\theta_{Ti} = A\theta_{Fi} + B, \quad a_{Tj} = \frac{a_{Fj}}{A}, \quad b_{Tj} = Ab_{Fj} + B, \quad c_{Tj} = c_{Fj}. \quad (2)$$

For the same item j and the same examinee i , the difference in the probability of getting a correct answer based on the two scales is expressed as

$$d_{ij} = p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Fi}; a_{Fj}, b_{Fj}, c_{Fj}) = p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}\left(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj}\right). \quad (3)$$

A loss function L evaluating the resultant losses, d_{ij} , is defined as

$$L(d_{ij}) = \sum_i \sum_j w_{ij} d_{ij}^2, \quad (4)$$

where w_{ij} is the weight assigned to the probability difference for item j and examinee i , and minimizing this equation gives the coefficients A and B . There are many ways to define the weight function, but, as explained in the previous section, the authors will focus on just two of them in this study: the Huber function and the LAV. Notice that the Haebara method is a special case of this method, when all weights are equal to 1.

Weighting with the area between ICCs. In the Area-Weighted method, the weights are defined using a Huber function:

$$w_{ij} = w_j = \begin{cases} 1 & |e_j| \leq k \\ k/|e_j| & |e_j| > k \end{cases}, \quad (5)$$

where k is the tuning constant and was set to 1.345 to obtain reasonable high efficiency (Holland & Welsch, 1977), and

$$e_j = \frac{\text{Area}_j}{\sigma_j}. \quad (6)$$

In Equation 6, Area_j is the area enclosed between two ICCs within $\theta = -4$ and $\theta = 4$ for item j (Linn et al., 1981; Rudner, 1977) and can be estimated by

$$\text{Area}_j = \sum_q |p_j(\theta_q; a_{Tj}, b_{Tj}, c_{Tj}) - p_j(\theta_q; a_{Fj}, b_{Fj}, c_{Fj})| \cdot \Delta\theta, \quad (7)$$

where q indexes the quadrature points and $\Delta\theta$ is the interval of abilities between two quadrature points. In Equation 6, σ_j is the standard deviation of Area_j and can be robustly estimated by

$$\frac{MAD(\text{Area}_j)}{0.6745},$$

where MAD is the Median Absolute Deviation (Wilcox, 2012). Because the weight function and the estimates of coefficients A and B depend on each other, the authors used an iterative procedure to minimize Equation 4. Given initial values of A and B , they computed the weight function. With the computed weight function, they obtained updated values of A and B . This procedure was repeated until Equation 4 was minimized.

Weighting with the LAV. Instead of minimizing the sum of squared differences of ICCs, the method of LAV minimizes the absolute difference between two ICCs of the equated test forms. The loss function is defined as

$$L_{LAV}(d_{ij}) = \sum_i \sum_j |d_{ij}|. \quad (8)$$

Note that this is equivalent to defining $w_{ij} = 1/|d_{ij}|$ in Equation 4. This method is robust in that a large value in d_{ij} is corresponding to a small weight for the squared difference.

Simulation Studies

The authors carried out two simulation studies to evaluate the two robust scale transformation methods. In the first study, they did not simulate outliers and expected the robust scale transformation methods would work as well as the Stocking-Lord method in terms of item parameter recovery. In the second study, they simulated outliers and compared the performance of the two robust scale transformation methods with the Stocking-Lord method.

Study 1: Recovery. In this study, the authors used the item parameter values presented by Kolen and Brennan (2004, p. 192) as the true item parameter values. There were 36 items in each test form, and 12 of them were internal common items. Setting the ability distribution for the group of examinees taking the old test form to $\theta \sim N(0, 1)$, five ability distributions, $\theta \sim N(0, 1)$, $\theta \sim N(0.25, 1.1^2)$, $\theta \sim N(-0.25, 1.1^2)$, $\theta \sim N(0.5, 1.2^2)$, and $\theta \sim N(-0.5, 1.2^2)$, were used for simulating examinees taking the new form. Given the ability distributions and the Kolen and Brennan item parameter values, the authors simulated item responses for 1,000 examinees using the 3PL IRT model. The simulated item responses were calibrated using the computer program BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2003). The item parameter estimates for the new form were transformed to the old form scale using the three scale transformation methods: the Stocking-Lord method, the Area-Weighted method, and the LAV method. The coefficients yielded by the Stocking-Lord method were used as initial values for the robust scale transformation methods.

To evaluate the recovery of the true scale transformation, the authors repeated the whole procedure 100 times and then calculated the mean, RMSE, and bias of the two scale transformation coefficients over the 100 replications. If ω denotes the true value of a scale transformation coefficient, either A or B , then the bias of the coefficient is defined by

$$\text{Bias}[\hat{\omega}] = E(\hat{\omega}) - \omega = \bar{\omega} - \omega, \quad (9)$$

where $\hat{\omega}$ represents an estimate of the true scale transformation coefficient, and $\bar{\omega}$ represents the mean of the scale transformation coefficient. The RMSE is defined by

$$\text{RMSE}[\hat{\omega}] = \sqrt{[SE(\hat{\omega})]^2 + [\text{Bias}(\hat{\omega})]^2}, \quad (10)$$

where

$$SE(\hat{\omega}) = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\omega}_r - \bar{\omega})^2}. \quad (11)$$

Equation 11 equals the standard error of estimate of the scale transformation coefficient over R replications. The true value of each scale transformation coefficient, either A or B , depends on the ability distribution used to generate the item responses. For example, the true values of the scale transformation coefficients should be $A = 1.1$ and $B = 0.25$ for $\theta \sim N(0.25, 1.1^2)$.

Study 2: Robustness to outlier. The same item parameter values as in the recovery study were used in this simulation study. Score conversions from IRT true score equating (Kolen & Brennan, 2004) using these item parameters were assumed to be the population equating relationship.

In most previous simulation studies evaluating the effects of common item outliers, only the b -parameter was manipulated (e.g., He et al., 2013; Wells, Subkoviak, & Serlin, 2002). In this study, the authors manipulated both a - and b -parameters so as to reflect the more typical situations encountered in practice. To simulate a simple outlier condition, the authors randomly selected a common item and adjusted both a - and b -parameter values according to different outlier conditions. Except for the *No Outlier* condition, a random number from a uniform distribution $\Delta a \sim U(0.1, 0.5)$ was used to adjust the a -parameter value (in a pilot study, the authors found that negative adjustments in a -parameter yielded similar results as positive adjustments). Similarly, the b -parameter value was simulated under four conditions representing mild to moderate outliers: $\Delta b \sim U(0.1, 0.5)$, $\Delta b \sim U(-0.5, -0.1)$, $\Delta b \sim U(0.5, 1.0)$, and $\Delta b \sim U(-1.0, -0.5)$. For example, if the original a - and b -parameter values were 1.1 and -0.1 , and the randomly generated numbers were 0.2 and -0.7 , then the manipulated a - and b -parameter values would be 1.3 and -0.8 . The simulated outlier was schematically presented in Figure 1. When simulating item responses for 1,000 examinees, the authors used the manipulated item parameter values for that one item. Three ability distributions, $\theta \sim N(0, 1)$, $\theta \sim N(0.25, 1.1^2)$, and $\theta \sim N(0.5, 1.2^2)$, were also used. Given the ability distributions and the item parameter values, 1,000 generated dichotomous item responses for the 36 test items served as the data for this part of the study.

The resulting item parameter estimates for the new form were put on the scale of the old form using the three scale transformation methods: the Stocking-Lord method, the Area-weighted method, and the LAV method. Other procedures were the same as those in Study 1. After scale transformation, IRT true score equating was conducted.

Similar to Study 1, the mean, RMSE, and bias of the two scale transformation coefficients were obtained in the presence of a simulated outlier. RMSE and bias of equated scores were used to evaluate the robustness of the proposed scale transformation methods in terms of the impact on equating results. For all score points combined, the authors also computed weighted absolute bias (WAB) and weighted root mean squared error (WRMSE) statistics.

Results

Study 1: Recovery

Table 1 shows the mean, RMSE, and bias of the scale transformation coefficients over 100 replications. The results indicate that the scale transformation coefficients were recovered well in all

Figure 1. Schematic presentation of a simulated outlier in the common item set. Left panel: Outlier condition (b), $\Delta b \sim U(0.1, 0.5)$; Right panel: Outlier condition (d), $\Delta b \sim U(0.5, 1.0)$.

studied conditions. All three scale transformation approaches produced small bias and RMSE. As can be seen, the differences among the approaches were negligible, which shows that the robust methods worked as well as the Stocking-Lord method when there was no outlier in the common-item set.

Study 2: Robustness to Outlier

Results of mean, RMSE, and bias of the scale transformation coefficients were reported only for the ability distribution of $N(0, 1)$ because the results for different ability distributions were similar. As shown in Table 2, all three scale transformation approaches produced large bias and RMSE values if an outlier was simulated. The Stocking-Lord method tended to yield larger RMSE values than the robust methods, which is more evident when a moderate outlier was simulated (i.e., the absolute change of b -parameter value was larger than 0.5). This finding suggests that the robust methods were more accurate in finding the scale transformation coefficients than the Stocking-Lord method when an outlier was present.

Table 3 summarizes the WAB results from using IRT equating with the three different transformation methods. When there was no simulated outlier, all methods yielded small WAB values (less than 0.030). The Stocking-Lord method under the condition of $\theta \sim N(0.25, 1.1^2)$ produced the largest WAB value of 0.027. This finding suggests that the newly proposed methods performed as well as the Stocking-Lord method in terms of bias when no outliers were present in the data. When an outlying common item was simulated, Table 3 shows that all methods tended to increase the WAB. The increase yielded by the Stocking-Lord method, however, was much larger than the two robust methods.

Table 4 shows the WRMSE results from doing an IRT equating after first applying the three scale transformation methods. As can be seen from this table, the robust scale transformation methods yielded larger WRMSE than the Stocking-Lord method under the condition of *No Outlier*, although the differences were small. When a common item outlier was simulated, Table 4 shows that the LAV method yielded the smallest WRMSE values under all outlier conditions and the Stocking-Lord method yielded the largest WRMSE values for most outlier conditions. These findings indicate that the proposed methods were more robust than the

Table 1. Mean, RMSE, and Bias of Scale Transformation Coefficients.

Ability distribution	Stocking-Lord		LAV		Area-weighted	
	A	B	A	B	A	B
$N(0, 1)$						
M	1.007	-0.008	1.006	-0.007	1.003	-0.008
RMSE	0.038	0.044	0.040	0.036	0.038	0.008
Bias	0.007	-0.008	0.006	-0.007	0.003	-0.008
$N(0.25, 1.1^2)$						
M	1.112	0.252	1.111	0.254	1.108	0.254
RMSE	0.045	0.039	0.051	0.042	0.048	0.040
Bias	0.012	0.002	0.011	0.004	0.008	0.004
$N(0.5, 1.2^2)$						
M	1.199	0.496	1.198	0.496	1.193	0.496
RMSE	0.037	0.041	0.045	0.045	0.045	0.042
Bias	-0.001	-0.004	-0.002	-0.004	-0.007	-0.004
$N(-0.25, 1.1^2)$						
M	1.100	-0.258	1.097	-0.254	1.096	-0.255
RMSE	0.038	0.042	0.039	0.048	0.041	0.043
Bias	0.000	-0.008	-0.003	-0.004	-0.004	-0.005
$N(-0.5, 1.2^2)$						
M	1.198	-0.495	1.198	-0.495	1.196	-0.495
RMSE	0.043	0.043	0.046	0.044	0.044	0.041
Bias	-0.002	0.005	-0.002	0.005	-0.004	0.005

Note. RMSE = root mean square errors; LAV = Least Absolute Values.

Table 2. Mean, RMSE, and Bias of Scale Transformation Coefficients Considering Simulated Outlier, $\theta \sim N(0, 1)$.

Outlier	Stocking-Lord		LAV		Area-weighted	
	A	B	A	B	A	B
No outlier						
M	1.007	-0.008	1.006	-0.007	1.003	-0.008
RMSE	0.038	0.044	0.040	0.036	0.038	0.008
Bias	0.007	-0.008	0.006	-0.007	0.003	-0.008
$\Delta b \sim U(0.1, 0.5)$						
M	1.018	-0.035	1.01	-0.007	1.011	-0.012
RMSE	0.065	0.064	0.053	0.055	0.055	0.014
Bias	0.018	-0.035	0.010	-0.007	0.011	-0.012
$\Delta b \sim U(-0.5, -0.1)$						
M	1.018	-0.031	1.01	-0.005	1.011	-0.011
RMSE	0.063	0.061	0.048	0.050	0.052	0.013
Bias	0.018	-0.031	0.010	-0.005	0.011	-0.011
$\Delta b \sim U(0.5, 1.0)$						
M	1.016	-0.065	1.015	-0.003	1.02	-0.008
RMSE	0.086	0.093	0.045	0.063	0.051	0.014
Bias	0.016	-0.065	0.015	-0.003	0.020	-0.008
$\Delta b \sim U(-1.0, -0.5)$						
M	1.013	-0.071	1.005	-0.006	1.011	-0.011
RMSE	0.091	0.091	0.046	0.054	0.050	0.016
Bias	0.013	-0.071	0.005	-0.006	0.011	-0.011

Note. RMSE = root mean square errors; LAV = Least Absolute Values.

Table 3. Weighted Absolute Bias for IRT True Score Equating.

Ability distribution	Outlier	Stockin-Lord	LAV	Area-weighted
$N(0, 1)$	a. No outlier	0.016	0.010	0.021
	b. $\Delta b \sim U(0.1, 0.5)$	0.126	0.032	0.016
	c. $\Delta b \sim U(-0.5, -0.1)$	0.143	0.019	0.045
	d. $\Delta b \sim U(0.5, 1.0)$	0.271	0.095	0.070
	e. $\Delta b \sim U(-1.0, -0.5)$	0.278	0.101	0.076
$N(0.25, 1.1^2)$	a. No outlier	0.027	0.020	0.014
	b. $\Delta b \sim U(0.1, 0.5)$	0.109	0.026	0.018
	c. $\Delta b \sim U(-0.5, -0.1)$	0.110	0.014	0.020
	d. $\Delta b \sim U(0.5, 1.0)$	0.278	0.072	0.057
	e. $\Delta b \sim U(-1.0, -0.5)$	0.315	0.062	0.041
$N(0.5, 1.2^2)$	a. No outlier	0.015	0.014	0.020
	b. $\Delta b \sim U(0.1, 0.5)$	0.131	0.023	0.013
	c. $\Delta b \sim U(-0.5, -0.1)$	0.130	0.023	0.014
	d. $\Delta b \sim U(0.5, 1.0)$	0.266	0.097	0.110
	e. $\Delta b \sim U(-1.0, -0.5)$	0.274	0.090	0.106

Note. Bold font type indicates the smallest value in each condition. IRT = item response theory; LAV = Least Absolute Values.

Table 4. Weighted RMSE for IRT True Score Equating.

Ability Distribution	Outlier	Stocking-Lord	LAV	Area-Weighted
$N(0, 1)$	a. No outlier	0.172	0.197	0.190
	b. $\Delta b \sim U(0.1, 0.5)$	0.253	0.229	0.259
	c. $\Delta b \sim U(-0.5, -0.1)$	0.273	0.226	0.267
	d. $\Delta b \sim U(0.5, 1.0)$	0.410	0.250	0.295
	e. $\Delta b \sim U(-1.0, -0.5)$	0.418	0.244	0.285
$N(0.25, 1.1^2)$	a. No outlier	0.161	0.195	0.191
	b. $\Delta b \sim U(0.1, 0.5)$	0.241	0.212	0.213
	c. $\Delta b \sim U(-0.5, -0.1)$	0.250	0.209	0.233
	d. $\Delta b \sim U(0.5, 1.0)$	0.439	0.219	0.280
	e. $\Delta b \sim U(-1.0, -0.5)$	0.478	0.241	0.274
$N(0.5, 1.2^2)$	a. No outlier	0.162	0.204	0.186
	b. $\Delta b \sim U(0.1, 0.5)$	0.248	0.231	0.251
	c. $\Delta b \sim U(-0.5, -0.1)$	0.248	0.227	0.247
	d. $\Delta b \sim U(0.5, 1.0)$	0.408	0.240	0.294
	e. $\Delta b \sim U(-1.0, -0.5)$	0.413	0.248	0.302

Note. Bold font type indicates the smallest value in each condition. RMSE = root mean square errors; IRT = item response theory; LAV = Least Absolute Values.

Stocking-Lord method when an outlier was present and performed as well as the Stocking-Lord method with no outliers.

When the authors evaluated bias and RMSE at each score point, they obtained results similar to the weighted statistics results. Although the shapes of the curves were slightly different for different ability distributions, there was little overall variability among them. In other words, the ability distribution did not have much impact on the scale transformation methods. As a consequence, only results for $\theta \sim N(0, 1)$ are presented.

Figure 2. The bias statistics of IRT equating, $\theta \sim N(0, 1)$: (a) no outlier; (b) $\Delta b \sim U(0.1, 0.5)$; (c) $\Delta b \sim U(-0.5, -0.1)$; (d) $\Delta b \sim U(0.5, 1.0)$; and (e) $\Delta b \sim U(-1.0, -0.5)$.
Note. IRT = item response theory; LAV = Least Absolute Values.

Figure 2 shows the bias of equated scores under different outlier conditions. As can be seen from this figure, all scale transformation methods yielded small, nearly identical bias values when no outlier was present. When a common item outlier was present, all methods tended to increase bias. The increase yielded by the Stocking-Lord method, however, was much larger than the other two methods. This finding is consistent with what was found in Table 3, that

Figure 3. The RMSE statistics of IRT equating, $\theta \sim N(0, 1)$: (a) no outlier; (b) $\Delta b \sim U(0.1, 0.5)$; (c) $\Delta b \sim U(-0.5, -0.1)$; (d) $\Delta b \sim U(0.5, 1.0)$; and (e) $\Delta b \sim U(-1.0, -0.5)$.
Note. RMSE = root mean square errors; IRT = item response theory; LAV = Least Absolute Values.

outliers affected equating accuracy, but the robust scale transformation methods were much less affected.

Figure 3 shows the RMSE of equated scores under various outlier conditions. As can be seen from this figure, the robust methods yielded slightly larger RMSE than the Stocking-Lord method at most score points when no outlier was simulated. The difference among these methods was small compared with the magnitude of RMSE. When an outlying common item was

simulated, Figure 3 shows that all methods tended to increase RMSE. The increase yielded by the Stocking-Lord method, however, was larger than the other two methods, especially when a moderate outlier was presented. Figure 3 also shows that the LAV method yielded the smallest RMSE at most score points when an outlier was present. These findings are consistent with what was found in Table 4 that the proposed scale transformation methods were more robust against outliers than the Stocking-Lord method. Between the proposed robust scale transformation methods, the LAV method tended to perform better in terms of total equating error.

Discussion and Conclusion

Results from this study show that the bias increased over fivefold under conditions of only mildly changed *b*-parameters (i.e., the absolute change was between 0.1 and 0.5). This finding confirms suggestions in Kolen and Brennan (2004) that practitioners need to examine and deal with outliers before conducting common item equating.

Previous research has focused on detecting common item outliers and then eliminating them from the test (He et al., 2013; Kolen & Brennan, 2004). Although eliminating outlying items can improve equating accuracy as compared with leaving the outlying items in the common item set, this practice has the potential to undermine content representativeness of the common item (Cook & Petersen, 1987). By contrast, the newly proposed robust scale transformation methods have the advantage of maintaining content representativeness and reducing the impact of outlying common item on the scale transformation and IRT true score equating under the studied conditions. This is accomplished through assigning smaller weights to potential outliers instead of eliminating them from the common item set. Eliminating an outlier is equivalent to assigning it a weight of 0, and including an outlier is equivalent to assigning it a weight of one. By contrast, the robust scale transformation methods assign weights according to the magnitude of outliers without being constrained by the dichotomy.

The results from this study show that the two robust scale transformation methods were effective in improving equating accuracy while maintaining content representativeness. Although the robust methods were slightly less accurate than the Stocking-Lord method when there were no outliers, they were considerably more accurate when there were outliers. The two proposed robust methods performed similarly in terms of equating bias, and the LAV method seemed to perform better than the Area-Weighted method in terms of RMSE. It should be noted that the results were obtained when only one outlier was simulated, although it is not uncommon to see more than one outlier in practice.

Equation 4 represents a general framework for scale transformation based on ICCs. As mentioned earlier, the Haebara (1980) method is a special case within this framework when all weights equal unity. Although not reported here, additional analyses indicated that the LAV method and the Area-Weighted method were also more robust than the Haebara method.

The present study could be extended in several ways. First, different weighting functions could be explored such as the Tukey bi-square function. Using different values for the tuning constant could also be explored. With the Area-Weighted method, the tuning constant governs the tradeoff between robustness and efficiency. A large value of the tuning constant corresponds to high efficiency but low robustness. Different values other than 1.345 could be considered. With a different tuning constant, the Area-Weighted method could outperform the LAV method.

Another extension of this study could be the inclusion of more than one common item outlier. This study considered only a single outlier because of the small number of common items in the Kolen and Brennan example. For examples with larger numbers of common items, the effects of multiple outliers could be studied. Multiple outliers could produce more complicated

effects. They could balance each other out producing little effect, or they might combine to produce even more inaccuracy. Future studies with multiple outliers should be considered. Nonetheless, the results from this study demonstrate that there is much to be gained by the use of robust methods when outliers are present and that their use results in little harm when they are not. Besides different numbers of common items, this study could also be extended to include different proportions of common items. The authors expect an outlier has larger impact on equating accuracy when the proportion of common items in a test is small compared with when the proportion is large. As a result, the superiority of the robust methods to the Stocking-Lord method will be more evident.

Authors' Note

An earlier version of this study was presented at the 79th Annual Meeting of the Psychometric Society, 2014. Much of the work for this study was conducted while the first author was a graduate student at the University of Missouri.

Acknowledgments

The authors thank Ze Wang and Christopher K. Wikle for discussions on this study. The authors also thank J. P. Kim, David J. Woodruff, Qing Yi, and anonymous reviewers for their review of 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.

References

- Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.
- Bejar, I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. 81-8). Princeton, NJ: Educational Testing Service.
- 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.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.
- 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.
- Holland, P., & Welsch, R. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics—Theory and Methods, 6*, 813-827.

- Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1, 799-821.
- 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. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.
- Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comparison. *Applied Psychological Measurement*, 5, 159-173.
- Lord, F. M. (1982). Item response theory and equating: A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Testing equating* (pp. 141 - 149). New York, NY: Academic Press.
- Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17, 179-193.
- Marco, G. L. (1977). Item characteristic curve solutions to the three intractable testing problems. *Journal of Educational Measurement*, 16, 139-160.
- Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement*, 25, 373-383.
- Rudner, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the annual meeting of the American Educational Research Association. New York, NY.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26, 77-87.
- Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis* (3rd ed.). New York, NY: Academic Press.
- 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.