# A Comparison of Trend Scoring and IRT Linking in Mixed-Format Test Equating

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA

Hua Wei
Lihua Yao, Defense Manpower Data Center

April 2013

ALWAYS LEARNING PEARSON

**About Pearson**
Everything we do at Pearson grows out of a clear mission: to help people make progress in their lives through personalized and connected learning solutions that are accessible, affordable, and that achieve results, focusing on college-and-career readiness, digital learning, educator effectiveness, and research for innovation and efficacy. Through our experience and expertise, investment in pioneering technologies, and promotion of collaboration throughout the education landscape, we continue to set the standard for leadership in education.For more information about Pearson, visit http://www.pearson.com/.

**About Pearson's Research & Innovation Network**
Our network **mission** is to spark innovation and create, connect, and communicate research and development that drives more effective learning. Our **vision** is students and educators learning in new ways so they can progress faster in a digital world**.** Pearson's research papers share our experts' perspectives with educators, researchers, policy makers and other stakeholders. Pearson's research publications may be obtained at: http://researchnetwork.pearson.com/.

**Abstract**

Characteristics of constructed-response (CR) items bring complications to the equating of mixed-format tests. Variations of rater severity across scoring cycles, if not adjusted for, become a potential source of errors in mixed-format equating. The trend scoring method that involves rescoring a set of anchor CR item responses from the previous test administration in the current scoring cycle has been implemented in practice and shown effective in disentangling the differences in rater severity from the differences in examinee group ability. However, the trend scoring method is costly and time-consuming. An alternative equating procedure that involves calibrating item response data with a rater model is compared to the trend scoring method in simulated conditions. Results show that the IRT linking method through the rater model outperformed the trend scoring method under different conditions of across-year rater severity variation, proportion of CR items in the anchor set, and group ability distributions. The rater model that accounts for rater severity variations while avoiding the need for rescoring anchor responses is worth consideration for operational use.

*Keywords:* mixed-format test, rater severity variation, trend scoring, rater model

A Comparison of Trend Scoring and IRT Linking in Mixed-Format Test Equating

## Introduction

Mixed-format tests demonstrate the strengths of multiple-choice and constructed-response items, and are more appropriate for measuring learning than single-format tests (Martinez, 1999). With the increasing use of mixed-format tests in large-scale assessment programs, finding ways to equate mixed-format tests has become a center of interest for psychometric research (e.g., Hagge, 2010; Kim & Lee, 2006). Equating mixed-format tests is more challenging than equating tests that are composed of multiple-choice (MC) items only, since characteristics of constructed-response (CR) items bring certain complications to the equating process (Walker, 2007). In the common-item nonequivalent groups design, the common item set should be a mix of both MC and CR items to be fully representative of the overall test. However, CR items are typically scored by human raters, and the scoring criteria may be applied more leniently or harshly by the raters than planned. In this sense, the common CR items are not really common because they are subject to changes in rater severity from year to year. Variations of rater severity, if not adjusted for, become a potential source of errors in mixed-format equating.

To separate out the across-year differences in rater severity from the differences in examinee group ability, Tate (1999, 2000) proposed a linking method that involves inserting a sample of anchor papers, called "trend papers", from the population in Year 1 into the rating process for Year 2, and having the raters from Year 2 rescore those anchor papers. This trend scoring method has shown to result in the smallest equating error, when compared with other equating procedures that do not involve trend scoring (Kim, Walker, & McHale, 2010a, 2010b; Tate, 2003).

However, rescoring trend papers adds to operational cost and scoring time. Alternatively, there are a number of IRT-based models that incorporate rater effects and

thus avoid the efforts of trend scoring. The multi-facet Rasch model (Linacre, 1989), also known as the FACETS model, is widely used for modeling rater effects. The FACETS model includes a rater effect by shifting the item response function (IRF) up or down the ability scale. The FACETS model for a polytomously scored item takes the form as:

$$p_{ij}(\theta_n) = \frac{\exp\left[\sum_{k=0}^{x_i}\left(\theta_n - b_i - d_k - r_j\right)\right]}{\sum_{w=0}^{M_i}\exp\left[\sum_{k=0}^{w}\left(\theta_n - b_i - d_k - r_j\right)\right]}$$

where $\theta_n$ is the ability measure for person *n*, $b_i$ is the item difficulty parameter for item *i*, $d_k$ is the threshold parameter for category *k*, and $r_j$ is the rater parameter for rater $j$. $r_j$ denotes the rater's tendency to be either harsh or lenient, and is modeled as constant across items. $d_k's$ are also modeled as constant across items, and, therefore, not appropriate for items with varying category threshold parameters.

Muraki's rater effect model (1993) generalizes the FACETS model by including a discrimination parameter that varies by item. It takes the following form:

$$p_{ijk}(\theta_n) = \frac{\exp\left[\sum_{k=0}^{x_i} Da_i\left(\theta_n - b_i - d_{ik} - r_j\right)\right]}{\sum_{w=0}^{M_i}\exp\left[\sum_{k=0}^{w} Da_i\left(\theta_n - b_i - d_{ik} - r_j\right)\right]}$$

where $a_i$ is the item discrimination parameter for item *i*, and $d_{ik}$ is the threshold parameter for category *k* of item *i*. However, Muraki's rater effect model can only handle response data to CR items that are polytomously scored, and, therefore, is not suitable for data from mixed-format tests.

The hierarchical rater model (HRM) introduced by Patz, Junker, Johnson, and Mariano (2002) improves the FACETS model and Muraki's rater effect model by accounting

for local dependence in the rating data, and provides more accurate measures of rater effects. However, the HRM is more suitable for the design in which each CR item is rated by a larger number of raters. Moreover, the HRM is designed to model rated responses and not ideal for handling data that arise from mixed-format tests.

The rater model developed by Yao (2010b) is an IRT-based rater model that can be widely generalized. It can incorporate multiple parameters, including item difficulty, discrimination, and rater, and one or multiple latent traits. In addition, its capability of handling MC and CR items at the same time makes it particularly applicable in the equating of mixed-format tests. A previous study has shown that Yao's rater model recovers the rater, item, and ability parameters very well in CR-only tests and tests that consist of a mix of MC and CR items (Wang & Yao, 2012). The potential of Yao's rater model in the equating of mixed-format tests is promising and merits investigation.

The purpose of this study was to compare the performance of the trend scoring method and the IRT linking method via the application of Yao's rater model in the equating of a mixed-format test, which followed the common-item nonequivalent groups design. Simulated data were used to explore how different conditions of across-year rater severity variations, anchor item composition, and group ability distributions might have impacted the equating results of the two methods.

## Method

### Description of Test Forms

In this study, two test forms, Forms X and Y, were simulated. Form Y was considered as the base form administered in Year 1 to the simulees of Group 1, and Form X as the new form administered in Year 2 to the simulees in Group 2. Each form consisted of 54 items including 48 MC items and 6 five-category (i.e., integer scores ranging from 0 to 4) CR items. Within each form, there were 36 unique items and 18 items that were common between the two forms.

## Study Conditions

The three factors that were of interest in this study are:

a. Variability in rater severity: In Yao's rater model, the rater parameter indicates the rater's degree of severity, and a larger value indicates a more severe rater. In this study, eleven raters were simulated with parameter values of -1.0, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, and 1.0. The CR response data for simulees in Year 1 were generated using the rater parameter values as above. The year-to-year change in rater severity was represented by a change in the rater parameter for each CR item. In this study, the rater severity decreased from Year 1 to Year 2, and the change was manipulated at two levels: 0.1 and 0.3. To put it more specifically, response data for CR items in Year 2 were generated in the same way, but by decreasing the rater parameters by 0.1 or 0.3.

b. MC to CR ratio in the anchor set: The ratio between MC and CR items in the anchor set was manipulated at two levels: 1) 16 MC and 2 CR items; and 2) 14 MC and 4 CR items.

c. Group ability distributions: The ability parameters for Group 1 were simulated from the standard normal distribution. The ability parameters for Group 2 were simulated from the same distribution under the equivalent condition, and from the distribution of $N(0.2,1)$ under the non-equivalent condition.

The three factors were fully crossed to produce a total of 8 conditions. Each condition was replicated 40 times.

## Data Generation

### Yao's Rater Model

Yao's rater model is an extension of the generalized two-parameter partial credit model with an additional rater parameter. For a polytomously scored item *j* ( $j = 1,...,J$ ) with rater $R_r$ ( $r = 1,...,M$ ), the probability of a response $k-1$ for a student with ability $\theta_i$ is given by the rater model as:

$$P_{ijkr} = P(x_{ijr} = k-1 | \theta_i, \vec{\beta}_j, R_r) = \frac{e^{(k-1)(\beta_j\theta_i - R_r) - \sum_{t=1}^{k}\beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{\left[(m-1)(\beta_j\theta_i - R_r) - \sum_{t=1}^{m}\beta_{\delta_t j}\right]}},$$

where $X_{ijr} = 0,...,K_j - 1$ is the response of student *i* to item *j*. $\beta_{\delta_k j}$ (k = 1, 2, . . . , $K_j$ ) is the threshold parameter, $\beta_{\delta_1 j} = 0$, and $K_j$ is the number of response categories for the *j*th item. The parameters for the *j*th item are $\vec{\beta}_j = (\beta_j, \beta_{\delta_2 j},..., \beta_{\delta_{K_j} j})$. It is clear that

$$P_{ij1r} = \frac{1}{\sum_{m=1}^{K_j} e^{\left[(m-1)(\beta_j\theta_i - R_r) - \sum_{t=1}^{m}\beta_{\delta_t j}\right]}},$$

$$P_{ijkr} = P_{ij(k-1)r} e^{\beta_j\theta_i - R_r - \beta_{\delta_k j}},$$

$$\log\frac{P_{ijkr}}{P_{ij(k-1)r}} = \beta_j\theta_i - R_r - \beta_{\delta_k j},$$

where $k = 2,...,K_j$.

Yao's rater model is the same as the regular generalized one-parameter partial credit model if $\beta_j = 1$ and $R_r = 0$. If $R_r = 0$, Yao's rater model is the same as the regular generalized two-parameter partial credit model. In this study, it is assumed that $\beta_j = 1$, and, therefore, the model is reduced to a generalization of the Rasch partial credit model

which incorporates a rater parameter. Item response data were generated based on Yao's rater model with the computer program BMIRTII (Yao, 2010a).

The item parameters used for generating response data were selected from a calibrated item pool with both MC and CR item types. Table 1 provides the averages of the b-parameters for the two test forms as well as the two anchor forms.

Table 1. Average item difficulty

| | Form Y | Form X | Anchor Form 1 MC/CR ratio =16:2 | Anchor Form 2 MC/CR ratio =14:4 |
|---|---|---|---|---|
| Average b-parameter | 0.0626 | 0.0637 | 0.0167 | -0.0233 |

As shown in Table 1, the two full forms were parallel in terms of average difficulty. The two anchor forms were similar to each other and the full forms in average difficulty.

**Generating item response data**

Under each simulated condition, two sets of item response data were generated independently, one for each form. Yao's rater model was used to generate the MC and CR item responses for the two groups of simulees in Year 1 and Year 2. A random assignment design was assumed for CR scoring, in which each of the eleven raters was assigned to rate a random sample of 500 responses to each CR item. To generate the response data for the CR portion of Form X for Group 2, the rater parameters were decreased by 0.1 or 0.3 to reflect changes in rater severity. Similarly, under the trend scoring approach, the rescored response data to each anchor CR item for the rescored sample from Group 1 were also generated by decreasing the rater parameters by 0.1 or 0.3.

**Equating Procedures**

In this study, three different approaches of equating were implemented with the simulated response data in Year 1 and Year 2. The equating procedures of the three methods are described as below:

**Baseline**

Under each simulated condition, a baseline equating approach that did not involve rescoring of anchor papers or any adjustment for rater effects was implemented. The baseline approach involved the general steps of calibrating Form X and placing the item parameters onto the scale of Form Y. The item parameters for the two forms were estimated separately in free calibrations by using the Rasch partial credit model in WINSTEPS, Version 3.64.2 (Linacre, 2006). The freely calibrated item parameters for Form Y were considered to be on the base scale, and the freely calibrated item parameters of Form X were equated to the base scale through the mean/sigma method. The means and standard deviations of the b-parameter estimates for the common items for the base form and the new form were entered into the following equations to find out the transformation constants:

$$slope = \frac{\sigma(b_{base})}{\sigma(b_{new})}$$

$$\text{intercept} = \mu(b_{base}) - slope \cdot \mu(b_{new})$$

The ability parameters for the simulees in Group 2 were equated to the base scale by applying the scale transformation constants through the following formula:

$$\theta_{transformed} = \text{intercept} + \text{slope} \cdot \bar{\theta}\,.$$

**Trend Scoring**

Compared with the baseline approach, the trend scoring approach includes an additional step of adjusting for across-year rater severity variations. The trend scoring method, in essence, defines a special linking study that involves selecting a representative sample of responses to each CR anchor item from Year 1 and having them rescored by the raters in Year 2. In other words, for each CR anchor item, this selected sample have two sets of scores, one assigned by raters from Year 1, and the other by raters from Year 2. In this study, a representative sample of 1,100 simulees was selected for rescoring, and the steps of adjusting for the across-year rater variability were as follows:

1. The freely calibrated item parameters for Form Y based on the responses of Group 1 were treated as the Year 1 estimates.

2. For the selected 1,100 simulees, the rescored response data and the Year 1 response data were merged together. The item parameters for the Year 2 rescored CR anchor items were calibrated by anchoring on the Year 1 estimated parameters for both MC and CR items. By doing this, the parameters for the two rescored CR items were placed onto the Year 1 scale.

3. For each anchor CR item, there were two $b_i$ estimates, one based on the responses scored by the Year 1 rater pool, and the other based on the rescored responses scored by the Year 2 rater pool. The difference between the two $b_i$ estimates was the rater-effect adjustment value for that particular CR item.

After the rater-effect adjustment value was calculated, this value was added to the Year 2 estimates of anchor CR items before proceeding to the step of calculating the scale transformation constants. Similarly, all the item calibrations under the trend scoring method were performed with WINSTEPS 3.64.2.

**Yao's Rater Model**

Yao's rater model adjusts for the rater effect by incorporating a rater parameter in the model, and thus avoids the efforts of rescoring anchor papers. Estimation of the rater model is done in BMIRTII, which implements the Markov chain Monte Carlo (MCMC) Metropolis Hastings algorithm. By adopting a perspective of Bayesian inference, the MCMC method imposes a prior distribution for each parameter in the model and estimates the full conditional posterior distribution of each parameter given the response data and other parameters in the model. The initial samples from the posterior distribution are discarded as burn-ins until the Markov chain has stabilized. After the burn-ins, a sample is taken from each parameter's posterior distribution and the sample mean is treated as the estimate of the parameter.

Information about the priors used for estimating the parameters in Yao's rater model is provided in Table 2.

Table 2. Priors in MCMC estimation of Yao's rater model

| Parameters | MC Items | CR Items |
|---|---|---|
| Ability | $N(0,1)$ | $N(0,1)$ |
| Item difficulty | $N(0,1.5)$ | N/A |
| Rater | $N(0,0.6)$ | $N(0,0.6)$ |
| Threshold | N/A | $N(0,1.5)$ |

In this study, 2000 iterations were discarded as burn-ins, and the parameter estimates were obtained from a sample of 5000 iterations after the burn-in. The item, ability, and rater parameters for Year 1 and Year 2 were estimated in separate runs. The rater parameter estimates and the item parameter estimates for the common items

obtained from the two separate calibrations were used to derive the transformation constants through the following formulae:

$$slope = \frac{sqrt\left(\text{var}(r_{base}) + \text{var}(b_{base})\right)}{sqrt\left(\text{var}(r_{new}) + \text{var}(b_{new})\right)}$$

$$\text{intercept } = \left(mean(r_{base}) + mean(b_{base})\right) - slope \cdot \left(mean(r_{new}) + mean(b_{new})\right)$$

By applying the transformation constants, the estimated ability parameters for the simulees in Group 2 from Year 2 could be placed onto the base scale.

**Evaluation Criteria**

Under each simulated condition, three sets of ability estimates were obtained for the simulees in Group 2, one from the baseline method, one from the trend scoring method and one from the rater model linking method. The ability estimates from each method were compared with the ability parameter values that were used to generate the item response data. The amount of difference was evaluated by the root mean squared error (RMSE), which is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i}^{I}\left(\frac{1}{H}\sum_{h}^{H}\left(\hat{\theta}_{ih} - \theta_{ih}\right)^2\right)}{I}},$$

where $\hat{\theta}_{ih}$ is the equated ability estimate for the *i*th simulee from one of the three equating approaches for the *h*th replication, and $\theta_{ih}$ is the "true" ability parameter value that was used for generating the response data.

Bias was another index calculated to evaluate the difference. It is expressed mathematically as:

$$Bias = \frac{\sum_{i}^{I}\left(\frac{1}{H}\sum_{h}^{H}\left(\hat{\theta}_{ih} - \theta_{ih}\right)\right)}{I}.$$

The equated ability estimates for simulees in Group 2 were also correlated with the "true" ability parameter values, and the average Pearson product-moment correlation coefficient was obtained across the 40 iterations for each method under each studied condition.

## Results

In this study, an equating scenario which follows the common-item nonequivalent groups design was simulated. Rater effects were simulated in the responses to CR items across the two years. Two different equating approaches were implemented that accounted for the rater effects in the process. A baseline approach which made no attempt in adjusting for rater effects was also implemented as a reference. The equated ability parameters for simulees in Group 2 from Year 2 were compared with the "true" ability values to evaluate the performance of each method. Table 3 summarizes the RMSE and Bias values over the 40 replications for each method under each studied condition.

Table 3. Root mean squared error (RMSE) and bias in equated ability parameters for Group 2 from Year 2

| Rater Severity Variation | MC:CR in Anchor Set | Group Ability Distributions | Baseline | | Trend scoring | | Rater model | |
|---|---|---|---|---|---|---|---|---|
| | | | RMSE | Bias | RMSE | Bias | RMSE | Bias |
| 0.1 | 16:2 | Equivalent | 0.2909 | 0.0081 | 0.2968 | -0.0273 | 0.2728 | 0.0320 |
| | | Non-equivalent | 0.2889 | 0.0031 | 0.2947 | -0.0288 | 0.2719 | 0.0354 |
| | 14:4 | Equivalent | 0.3124 | -0.0096 | 0.3372 | -0.0756 | 0.2752 | 0.0493 |
| | | Non-equivalent | 0.3087 | -0.0061 | 0.3314 | -0.0661 | 0.2748 | 0.0531 |
| 0.3 | 16:2 | Equivalent | 0.3048 | 0.0631 | 0.2951 | -0.0275 | 0.2877 | 0.0937 |
| | | Non-equivalent | 0.3053 | 0.0653 | 0.2945 | -0.0280 | 0.2915 | 0.1060 |
| | 14:4 | Equivalent | 0.3443 | 0.0851 | 0.3345 | -0.0772 | 0.3119 | 0.1509 |
| | | Non-equivalent | 0.3470 | 0.0986 | 0.3303 | -0.0661 | 0.3179 | 0.1634 |

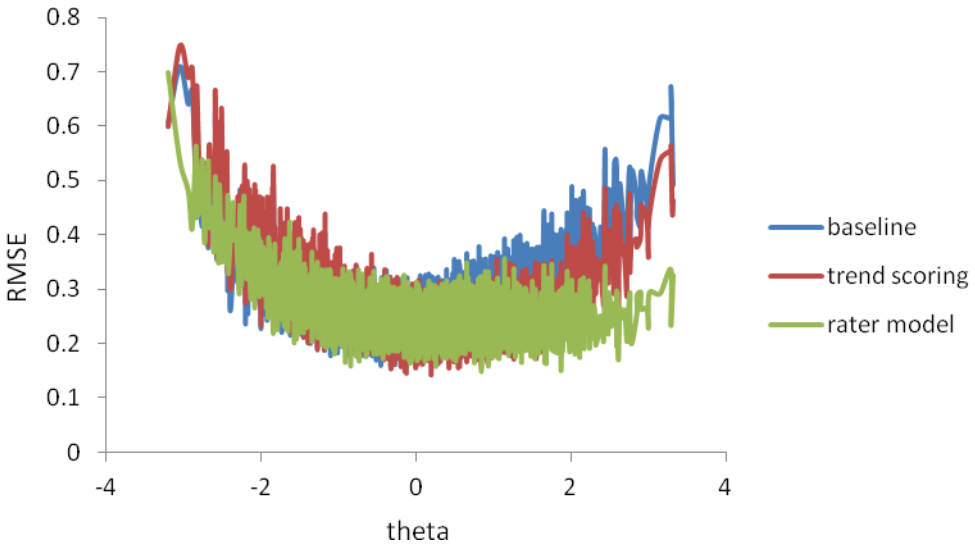
As shown in the table, the rater model was associated with smaller RMSE values under all the simulated conditions, which suggested that the rater model tended to produce more accurate ability parameter estimates in the presence of rater effects. Although the differences were not substantial, the pattern was clear. As the between-year rater severity difference increased from 0.1 to 0.3 logit, the values of RMSE increased for the baseline and rater model methods. In contrast, the trend scoring method seemed to be insensitive to the rater severity changes, if other conditions were held constant. For all the three methods, the RMSE values increased when the number of CR items in the anchor set increased from 2 to 4. This is understandable because rater effects only reside in CR items. The more CR items in the anchor set, the greater the impact rater effects bring on the equating results. In contrast to the other factors, the factor of group ability distributions did not seem to have an obvious effect on the RMSE values. The RMSE values were the largest under the conditions with rater severity difference equal to 0.3 and 4 CR items in the anchor set.

Similar to what was seen among the RMSE values, the Bias values were sensitive to the changes in rater severity. A greater degree of rater severity variation between years was associated with larger values of Bias for the baseline and rater model methods. The Bias values for the trend scoring method were similar between the two conditions of rater severity variation. A larger proportion of CR items in the anchor set was associated with larger Bias values for all the three methods. Again, the factor of group ability distribution did not have any influence on the Bias values for all the three methods. However, unlike the conclusion drawn for the RMSE, the absolute Bias values were the largest for the rater model under all the studied conditions. This may be explained by the Bayesian estimation algorithm of the rater model, which tends to produce overly biased estimates at the two ends of the ability scale.

RMSE and Bias values conditional on the "true" ability levels were also obtained for each method and compared among the three methods. Graph 1 shows the conditional RMSE

values under the simulated condition with rater severity variation equal to 0.3, 4 CR items in the anchor set, and equivalent ability distributions between years.

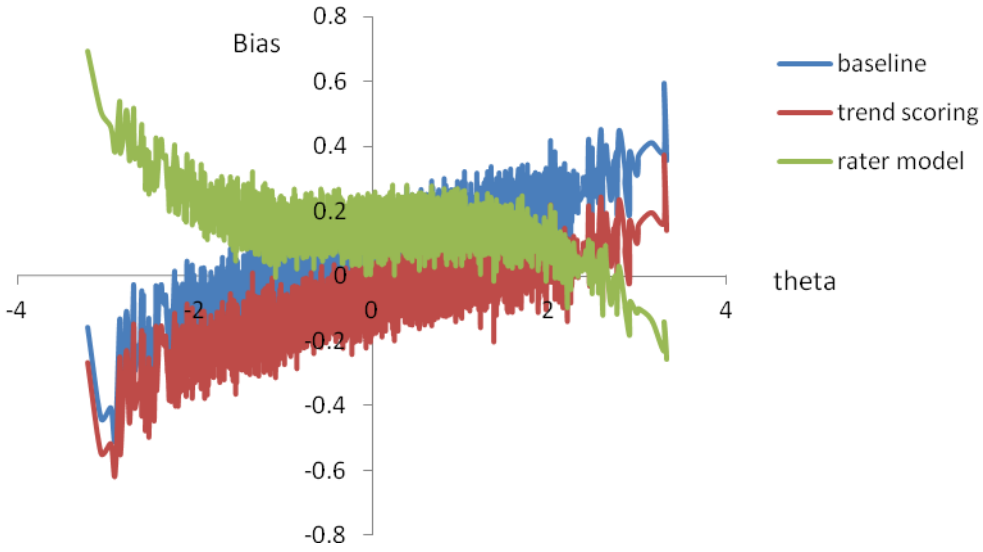
Graph 1. A comparison of conditional RMSEs among the three equating methods under one simulated condition




A comparison of the conditional RMSE values shows that the three methods behaved similarly in the middle of the ability scale. The rater model led to smaller RMSE values than the other two methods across the entire span of the continuum, except at the lower extreme of the ability scale. The baseline and trend scoring methods produced conditional RMSE values that shared some overlapping, but the trend scoring method behaved better with theta values above 0. All these findings are in accordance with what is presented in Table 3.

Graph 2 shows the conditional Bias values under the same simulated condition (i.e., rater severity variation equal to 0.3, 4 CR items in the anchor set, and equivalent ability distributions between years).

Graph 2. A comparison of conditional Bias values among the three equating methods under one simulated condition




Compared with the baseline and trend scoring methods, the rater model was associated with smaller absolute Bias values across a wide range of theta values. However, the absolute Bias values were larger for the rater model at the lower end of the ability scale. This is not a surprise, considering that Bayesian estimates are especially biased at the two extremes of the latent continuum.

An ANOVA test was conducted to investigate the main effects and interaction effects of the three simulated factors and the factor of equating methods on the values of RMSE. Table 4 provides the results of the ANOVA test on RMSE.

Table 4. ANOVA test of the effects of the factors on RMSE

| Source | df | Mean Square | F | Sig. |
|---|---|---|---|---|
| method | 2 | 0.0702 | 2210.26 | < 0.0001 |
| rater_variation | 1 | 0.0724 | 2282.00 | < 0.0001 |
| CR_ratio | 1 | 0.1819 | 5730.48 | < 0.0001 |
| group_ability | 1 | 0.0001 | 2.32 | 0.1277 |
| method*rater_variation | 2 | 0.0216 | 680.05 | < 0.0001 |
| method*CR_ratio | 2 | 0.0121 | 382.28 | < 0.0001 |
| rater_variation*CR_ratio | 1 | 0.0115 | 362.40 | < 0.0001 |
| method*group_ability | 2 | 0.0006 | 17.68 | < 0.0001 |
| rater_variation*group_ability | 1 | 0.0009 | 28.00 | < 0.0001 |
| CR_ratio*group_ability | 1 | 0.0000 | 0.96 | 0.3279 |
| method*rater_variation*CR_ratio | 2 | 0.0033 | 105.42 | < 0.0001 |
| method*rater_variation*group_ability | 2 | 0.0001 | 2.73 | 0.0656 |
| method*CR_ratio*group_ability | 2 | 0.0001 | 4.31 | 0.0137 |
| rater_variation*CR_ratio*group_ability | 1 | 0.0001 | 1.69 | 0.1938 |
| method*rater_variation* CR_ratio*group_ability | 2 | 0.0000 | 0.55 | 0.5745 |

As shown in the table, each of the two simulated factors, rater severity variation and proportion of CR items in the anchor set, was found to have a statistically significant effect on the values of RMSE. The factor of group ability distribution did not have a significant main effect on RMSE values. The RMSE values also varied significantly across the three equating methods. The results of the ANOVA test confirmed the observed patterns in Table 3.

The equated ability parameters for the simulees in Group 2 for each of the three equating methods were also evaluated against the "true" ability parameters. Table 5 presents the Pearson product-moment correlation coefficients for each method under all the simulated conditions.

Table 5. Correlation coefficients between equated ability parameters and "true" ability parameters for Group 2 from Year 2

| Rater Severity Variation | MC:CR in Anchor Set | Group Ability Distributions | Baseline | Trend scoring | Rater model |
|---|---|---|---|---|---|
| 0.1 | 16:2 | Equivalent | 0.9582 | 0.9582 | 0.9627 |
| | | Non-equivalent | 0.9586 | 0.9586 | 0.9631 |
| | 14:4 | Equivalent | 0.9582 | 0.9582 | 0.9627 |
| | | Non-equivalent | 0.9586 | 0.9586 | 0.9631 |
| 0.3 | 16:2 | Equivalent | 0.9585 | 0.9585 | 0.9631 |
| | | Non-equivalent | 0.9586 | 0.9586 | 0.9633 |
| | 14:4 | Equivalent | 0.9585 | 0.9585 | 0.9631 |
| | | Non-equivalent | 0.9586 | 0.9586 | 0.9633 |

The correlation coefficients were above 0.95 for all three methods under all the studied conditions. The correlation coefficients were greater for the rater model than for the other two methods. In other words, the IRT linking approach through Yao's rater model produced equated ability estimates that approximated the "true" ability parameters more closely than the other two equating approaches. The equated ability parameters under the baseline and the trend scoring methods were reported to have identical correlations with the "true" parameters. This is explainable because the two methods started with calibrating item response data with the same model. What distinguished the two methods was the process of deriving the transformation constants, which only linearly transformed the

originally calibrated ability parameters but did not affect the correlations with the "true" ability parameters.

In addition, we calculated the Pearson product-moment correlation coefficients between the equated ability parameters for the rater model and the trend scoring method. Table 6 presents the correlation coefficients under all the simulated conditions.

Table 6. Correlation coefficients between equated ability parameters for the rater model and the trend scoring method

| Rater Severity Variation | MC:CR in Anchor Set | Group Ability Distributions | Trend scoring vs. Rater model |
|---|---|---|---|
| 0.1 | 16:2 | Equivalent | 0.9953 |
| | | Non-equivalent | 0.9952 |
| | 14:4 | Equivalent | 0.9953 |
| | | Non-equivalent | 0.9952 |
| 0.3 | 16:2 | Equivalent | 0.9952 |
| | | Non-equivalent | 0.9950 |
| | 14:4 | Equivalent | 0.9952 |
| | | Non-equivalent | 0.9950 |

As shown in the table, the equated ability parameters obtained from the trend scoring approach correlated almost perfectly with the equated ability parameters from the rater model under all the studied conditions. This finding suggests that Yao's rater model is a potential alternative to the trend scoring method in the equating of mixed-format tests, if the purpose of the test is to rank order examinees by using their ability estimates.

## Conclusion

In this study, an equating scenario, in the context of the common-item nonequivalent groups design, was simulated that involved item response data from two consecutive years for a mixed-format test. Rater severity was simulated to vary between the two years in the scoring of constructed response items. Equating of the new form and the ability parameters for the simulees from Year 2 was implemented through each of three methods. The baseline equating approach represented the standard equating methodology that did not make any attempt to account for rater effects. The trend scoring method had the two groups of raters rate a common set of responses to the anchor CR items and disentangled rater severity difference from true ability difference. Yao's rater model explicitly modeled the rater effect by having it as a separate term in the model form, and thus separated it out from true ability. Comparison of the three equating methods was made under simulated conditions with different degrees of across-year rater severity variation, proportion of CR items in the anchor set, and group ability distribution.

Equated ability parameters for the simulees from Year 2 were evaluated against the ability values that were used for generating the response data. The rater model outperformed the other two methods in the sense that it produced smaller RMSE values and higher correlations with the "true" values under all the simulated conditions. A comparison of the conditional RMSE and Bias values showed that the rater model behaved better than the other two methods across the entire range of the ability scale, except at the lower extreme of the continuum. Results from the ANOVA test also confirmed that the choice of equating method did have a significant impact on the resulting RMSE values.

Compared with the baseline method, the trend scoring method produced smaller RMSE and Bias values only under the conditions with a larger across-year rater severity difference. This is saying that the trend scoring method is only effective in adjusting for

larger amounts of rater effects. By contrast, the advantage of the rater model over the baseline method is obvious under all the simulated conditions with different values of rater severity variation.

With the advances of the common core and college and career initiatives that promote the expanded use of performance-based tasks, mixed-format tests will be increasingly used in educational assessment programs. Compared with the equating method that makes no adjustment for rater effects, the rater model and the trend scoring method tend to produce more accurate equating results. Results of the study show that the advantages of these two methods are more pronounced under the conditions with greater across-year rater severity differences and larger proportions of CR items in the anchor form. Moreover, the errors of ignoring rater effects in the equating of mixed-format tests are likely to be accumulated over years, and the long-term impact of correcting for rater effects is yet to be determined by using longitudinal data.

Findings of the study have some practical implications in the field of mixed-format test equating. The trend scoring method has emerged as an effective solution to account for rater severity changes over time. However, it is expensive and difficult to implement in practice. The needs to retrieve "trend papers" from the previous administration and have them interspersed with papers in the current administration add significantly to scoring time and cost. In testing programs with a tight budget or a short turnaround, implementing the trend scoring method will be a challenge. Therefore, alternative methods, like Yao's rater model, that account for rater effects while avoiding the need for rescoring anchor papers are ideal for operational use.

**References**

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups.* Unpublished doctoral dissertation. Iowa City, IA: University of Iowa.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed format tests. *Journal of Educational Measurement, 43,* 53-76.

Kim, S., Walker, M.E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47,* 36-53.

Kim, S., Walker, M.E., & McHale, F. (2010b). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments*. Journal of Educational Measurement, 47,* 186-201.

Linacre, J. M. (1989). *Many-faceted Rasch Measurement.* Chicago, IL: MESA Press.

Linacre, J. M. (2006). *WINSTEPS version 3.60* [Computer Software]. Chicago, IL: Author.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34,* 207-218.

Muraki, E. (1993). *Variations of polytomous item response models: Raters' effect model, DIF model, and trend model.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36,* 336-346.

Tate, R.L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items*. Journal of Educational Measurement, 37,* 329–346.

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and Psychological Measurement, 63,* 893–914.

Walker, M. (2007, April). *Criteria to consider when reporting constructed-response scores.* Paper presented at the 2007 annual meeting of the National Council on Measurement and Evaluation, Chicago, IL.

Wang, Z., & Yao, L. (2012, April). *The effects of rater assignment and rater's severity on students' ability estimation for constructed-response items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Yao, L. (2010a). *BMIRTII: Bayesian Multivariate Item Response Theory* [Computer Program]. Monterey, CA: Defense Manpower Data Center.

Yao, L. (2010b). *Rater effect model.* Personal communication memo.