Listening. Learning. Leading.®

Research Report
ETS RR–13-23

The Effects of Rater Severity and Rater Distribution on Examinees' Ability Estimation for Constructed-Response Items

Zhen Wang

Lihua Yao

November 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

The Effects of Rater Severity and Rater Distribution on Examinees' Ability Estimation for Constructed-Response Items

Zhen Wang
ETS, Princeton, New Jersey

Lihua Yao
Defense Manpower Data Center, Seaside, California

November 2013

Action Editor: Donald E. Powers

Reviewers: Alina von Davier and Shelby J. Haberman

Abstract

The current study used simulated data to investigate the properties of a newly proposed method (Yao's rater model) for modeling rater severity and its distribution under different conditions. Our study examined the effects of rater severity, distributions of rater severity, the difference between item response theory (IRT) models with rater effect and without rater effect, and the difference between the precision of the ability estimates for tests composed of only constructed-response (CR) items and for tests composed of multiple-choice (MC) and CR items combined. Our results indicate that rater severity and its distribution can increase the bias of examinees' ability estimates and lower test reliability. Moreover, using an IRT model with rater effects can substantially increase the precision in the examinees' ability estimates, especially when the test was composed of only CR items. We also compared Yao's rater model with Muraki's rater effect model (1993) in terms of ability estimation accuracy and rater parameter recovery. The estimation results from Yao's rater model using Markov chain Monte Carlo (MCMC) were better than those from Muraki's rater effect model using marginal maximum likelihood.

Key words: IRT-based rater model, IRT**,** distributions of rater severity, MCMC

Acknowledgments

The authors are grateful to Alina von Davier, Shelby Haberman, and Don Powers for their advice during the review of the paper. The authors also thank Kim Fryer and Ruth Greenwood for their editorial comments and suggestions on earlier versions of the manuscript. The opinions expressed in this paper are those of the authors and not necessarily of ETS, the U.S. Department of Defense, or the United States government.

Constructed-response (CR) questions have been increasingly used in standardized assessments. CR items may include open-ended questions, structured performance tasks, and other kinds of free-response assessments that require the examinee to display certain skills and knowledge. Many testing programs use CR items in addition to multiple-choice (MC) items. However, there is strong evidence that human raters differ in their overall severity and in their evaluations of specific responses (Longford, 1995). Rater severity effects have been included in some measurement models and studied by DeCarlo, Kim, and Johnson (2011); Donoghue, McClellan, and Gladkova (2006); Engelhard (2002); Longford (1995); Patz, Junker, Johnson, and Mariano (2002); Wilson and Hoskens (2001); and Wolfe and Myford (1997). The results from their studies indicate that the bias or systematic error may be caused by varying degrees of rater leniency or strictness. If rater severity effects are not included in the measurement model, then the overall reliability will decrease (Donoghue et al., 2006).

There are a few research studies incorporating rater severity effects into the *item response theory* (IRT) models (Donoghue et al., 2006; Engelhard, 1996; Patz, 1997; Patz & Junker, 1999; Wilson & Hoskens, 2001). The *FACETS model* (Linacre, 1991) is the IRT model that allows for the estimation of differences in severity between raters, and thus eliminates rater bias from the estimates of the items and examinees' ability. However, FACETS has obvious problems with asymptotic bias, due to its use of the *joint (unconditional) maximum likelihood method*: "Its weaknesses are degree of statistical inconsistency under artificial conditions, and statistical bias with some very small data sets and certain idiosyncratic data configurations" (de Jong & Linacre, 1993, pp. 296–297). Although there are possible mitigations of the problem under certain scenarios involving many examinees, prompts, and raters by use of techniques proposed by Haberman (1977), the solution to the problem of joint estimation generally involves the use of *marginal maximum likelihood* or *conditional maximum likelihood.* Other IRT models are also available for modeling rater effects, such as *Muraki's rater effect models* (1993), the *hierarchical rater model* (HRM model; Patz, 1997), and the *rater bundle model* (Wilson & Hoskens, 2001). Patz's HRM model is a recent development that "accounts for marginal dependence between ratings of the same examinee's work" (Patz, 1997, p. 343), and it is more feasible for large numbers of ratings per item, where the FACETS model assumed independent relationship between such ratings. Wilson and Hoskens' rater bundle model also explicitly models the

dependence between multiple readings of the same examinee's work. However, the rater bundle model only works well for modeling a few specific dependencies.

In addition to rater severity effects, the distributions of rater severity (e.g., the manner in which raters are assigned to essays and examinees) can impact the accuracy of examinees' ability estimation. So far, only a few research studies have been conducted to investigate the effects of distributions of rater severity patterns on examinees' ability estimation (Hombo, Donoghue, & Thayer, 2001; Sykes, Ito, & Wang, 2008). Sykes et al. (2008) concluded that "rater bias on an examinee's set of responses may be minimized with the use of multiple readers though fewer than the number of items" (p. 47). Hombo et al. (2001) investigated different scoring designs and their impact on the accuracy of examinee ability estimation.

As a result of the new technology, the scoring and distributions of rater severity have changed dramatically for most projects; for instance, some raters can sign up for rating examinees' responses at home by using a central scoring system. This is very different from the traditional scoring process model, where each rater is assigned equal number of papers that are randomly selected and all the raters stay in the same scoring room. Some biased raters (e.g., those judged to be harsh or lenient) or those who tend to use middle scores may sign up more examinees for a particular assessment than other raters do, because they have time available, which may result in a nontrivial rater-bias effect. Such rater effects may have a serious impact. Raters often demonstrate consistent individual differences in their ratings, although sufficient training might reduce rater variability; incorporating rater effects into measurement models can increase the measurement precision if raters exhibit acceptable model-data fit. However, almost all testing companies choose to use IRT models with no rater severity parameter, due to the lack of a well-fitted rater model and software, as well as the lack of related research regarding rater severity effects and the distributions of rater severity.

The purpose of this study is to investigate the properties of a newly proposed method, Yao's rater model, for modeling rater severity under different conditions. Yao's rater model is an extension and improvement of the current existing models and software that implement Muraki's rater effect model and Linacre's Rasch-based rater model (1989). In the current study, we tested Yao's IRT-based rater model that handles both MC and CR items and uses simulated data to explore parameter recovery and rater distribution issues. We also used simulated data on CR-

only items and compared Yao's rater model with Muraki's rater effect model in terms of examinees' ability estimation accuracy and rater parameter recovery.

In the beginning of this paper, we provide an introduction, research purposes, and methods sections. In the methods section, we introduce two IRT-based rater models first. Then we describe the simulated data and four rater distribution patterns used and present the evaluation criteria for the rater model comparisons. Finally, we present results, conclusions, and discussions.

Research Purpose

The purpose of this research is:

1. to explore parameter recovery (raters, items, and examinees) using Yao's rater model (2012) for CR-only and MC plus CR combined tests and to compare Yao's rater model without rater parameters with Yao's model incorporating rater parameters;
2. to study the effects of four different rater distributions and rater severity on examinees' ability estimates, using a new IRT-based rater model (Yao, 2012); and
3. to compare two rater models, Muraki's rater effect model (1993) and Yao's rater model (2012), in terms of examinees' ability estimation.

Methods

Muraki's Rater Effect Model

Muraki's rater effect model (1993) in PARSCALE (Muraki & Bock, 1999) generalizes the FACETS model by adding a discrimination parameter (*a*-parameter) to the item. The estimation method used is the estimation and maximization (EM) algorithm for solving marginal likelihood equations. Muraki's rater effect model is expressed as follows (Barr & Raju, 2001):

$$P_{jk}(\theta_i) = \frac{\exp \sum_{m=0}^{x_j} Da_j(\theta_i - b_{jf} + \rho_r)}{\sum_{k=0}^{m_j} \exp \sum_{f=0}^{k} Da_j(\theta_i - b_{jf} + \rho_r)}, \quad (1)$$

where $x = 0,1, \ldots, m_j$ and $k = 0, 1, \ldots, m_j$. P_{jk} is the probability of rater r assigning examinee i to category k over category (k-1) in a polytomous item j. D is constant. θ_i is an examinee's (i's) proficiency, and a_j and b_{jf} are the item parameters for item j. In the case of a polytomous item

with *m* categories there will be one *a* parameter and *(m-1) b* parameters (Muraki & Bock, 1999). The tendency of the rater toward leniency or severity is separately accounted for in Muraki's rater effect model, denoted by ρ_r . The rater effect is modeled as constant across items.

Yao's Rater Model

The rater model developed by Yao (2012) is an IRT-based rater model. It can incorporate multiple parameters (item difficulty, discrimination, and rater) and one latent trait or multiple latent traits. Additionally it can handle both MC and CR items simultaneously. Yao's rater model was based on the Markov chain Monte Carlo estimation in BMIRTII software (Yao, 2010), using the Metropolis-Hastings algorithm. The procedure obtains samples from the posterior distribution—a product of the likelihood function and the priors. Because the Markov chain is not stable, initial samples often may not be valid. For each of the to-be-estimated parameters (raters, items, and abilities), the average of the Markov chain Monte Carlo sampling after the chain has reached its stationary or after the burn-in is the final estimate for that parameter.

Markov chain Monte Carlo methods have been widely used in every aspect of scientific inquiry, such as computational physics, biology, and linguistics. Using Markov chain Monte Carlo in Yao's rater model, we have the flexibility to extend the existing raters models (e.g., FACETS; PARSCALE) from one latent trait to multiple latent traits, and Markov chain Monte Carlo can handle both MC and CR items. Most rater models from the literature only handle tests that are composed of CR items. Since most tests are composed of both MC and CR items, we think it is important to include MC items along with CR items to compare with CR-only tests when estimating examinees' ability. Even though Yao's rater model can be multidimensional, only the unidimensional model is described here, as this study only used the unidimensional model.

Suppose there are *N* examinees, *J* items, and *M* raters with a continuous parameter *Rr* (a severity/leniency parameter) in the range of $(-\infty,+\infty)$, where r = 1, …, *M*. For a polytomously scored item *j*, *j* = 1, …, *J*, by rater *r*, the generalized two-parameter partial credit model provides the probability of a response $k-1$ for an examinee with ability θ_i:

$$P_{ijkr} = P(x_{ijr} = k-1|\theta_i, \vec{\beta}_j, R_r) = \frac{e^{(k-1)(\beta_j\theta_i - R_r) - \sum_{t=1}^{k}\beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{\left[(m-1)(\beta_j\theta_i - R_r) - \sum_{t=1}^{m}\beta_{\delta_t j}\right]}}, \quad (2)$$

where $Xijr = 0, \ldots, Kj - 1$ is the score of examinee i on item j. βj is the discrimination, $\beta_{\delta kj}$ for $k = 1, 2, \ldots, Kj$ are the threshold parameters, $\beta_{\delta 1j} = 0$, and Kj is the number of response categories for the jth item. The discrimination parameters for the jth item are $\beta j = (\beta j , \beta\delta_2 j , \ldots, \beta\delta_{Kj} j)$. It is clear that

$$P_{ij1r} = \frac{1}{\sum_{m=1}^{K_j} e^{\left[(m-1)(\beta_j\theta_i - R_r) - \sum_{t=1}^{m}\beta_{\delta_t j}\right]}}, \tag{3}$$

$$P_{ijkr} = P_{ij(k-1)r} e^{\beta_j\theta_i - R_r - \beta_{\delta_k j}}, \tag{4}$$

$$\log\frac{P_{ijkr}}{P_{ij(k-1)r}} = \beta_j\theta_i - R_r - \beta_{\delta_k j}, \tag{5}$$

where $k = 2, \ldots, Kj$. The rater parameters are $\vec{R} = (R1, \ldots, RM)$.

Yao's rater model is the same as the regular generalized one-parameter partial credit model, if $\beta j = 1$ and $Rr = 0$. If $Rr = 0$, Yao's rater model is the same as the regular generalized two-parameter partial credit model. The item discrimination parameter β_j is the same as Da in equation (1) and the threshold parameters are the same as that in equation (1) times Da. In this study, we used same scale ($D = 1$, $a = 1$) for both Yao's rater model and Muraki's rater effect model.

For MC items in BMIRTII, the models are the three-parameter logistic model. The multidimensional extension of the generalized two-parameter partial-credit model was described in Yao and Schwarz (2006). The multidimensional versions of the rater model are a similar extension of equation (2). When $R_r = 0$ for all raters, the multidimensional rater model is the same as the multidimensional generalized two-parameter partial model.

In the current study, we tested Yao's (2012) new-rater model using the basic model: a one latent trait and one-parameter IRT-based rater model. We used 10 data sets simulated from Yao's rater model itself to explore parameter recovery and rater distribution issues from the scoring process. Markov chain Monte Carlo was used to estimate item, ability, and rater-effect parameters based on Yao's rater model through running BMIRTII software (Yao, 2010). Comparison was done between CR-only tests and MC plus CR combined tests in terms of examinee ability estimates.

Simulation

Data Generation

We chose simulation methods to study the effects on examinees' ability estimation using different rater distribution patterns, and explored the use of Yao's rater model based on the simulated data. The simulated data allowed us to compare the estimated values from the different models with the true values (e.g., item parameters, rater parameters, examinee ability) so that different models could be evaluated meaningfully. Data were simulated incorporating rater effects under the one-parameter logistic (1PL) model using Yao's program, BMIRTII (Yao, 2010), which is equivalent to the two-parameter partial credit model with the item-discrimination (slope) parameters set to 1.0 for each item.

The data set created for the simulation was based on a set of item parameters calibrated from real assessment data, including 45 MC items and 15 CR items: six items with three categories (0, 1, 2), six items with four categories (0, 1, 2, 3), and three items with five categories (0, 1, 2, 3, 4). Due to time constraint, we used 10 replications with different seeds for each condition. Varying conditions are listed as the following:

1. True examinee ability: A simulation of 23,760 instances of examinee ability were created by selecting their ability values randomly from a normal (0,1) distribution;
2. Rater parameters: Eleven raters were chosen for the simulation with values −1.0, −0.7, −0.5, −0.2, −0.1, 0, 0.1, 0.2, 0.5, 0.7, and 1.0, where a larger value indicates a more severe rater. The size of the rater effects were chosen based on the analysis of a state assessment data.

Rater Distributions

The existence of rater effects in performance assessment data is often complicated by the distributions of rater severity. Four types of rater distributions were compared (see Table 1). In this study, all of an examinee's responses were scored by the same rater, and the distributions of examinees to raters were random.

Table 1

Responses Read by Each Rater in Rater Distributions 1–4

Distribution	Rater										
	1	2	3	4	5	6	7	8	9	10	11
1	2,160	2,160	2,160	2,160	2,160	2,160	2,160	2,160	2,160	2,160	2,160
2	4,000	4,000	4,000	1,470	1,470	1,470	1,470	1,470	1,470	1,470	1,470
3	1,470	1,470	1,470	1,470	1,470	1,470	1,470	1,470	4,000	4,000	4,000
4	1,470	1,470	1,470	1,470	4,000	4,000	4,000	1,470	1,470	1,470	1,470

The following is the description for each rater distribution pattern to address our research questions:

Rater Distribution 1 (D1)

In this rater distribution pattern, for each replication, each of the 11 raters was randomly assigned to read about a group of 2,160 examinees on each of the 15 CR items. The data of this rater distribution was analyzed to serve as a baseline. To simplify the rater distribution, we did not include multiple raters for one examinee's responses, although it is feasible.

Rater Distribution 2 (D2)

Three lenient raters (−1.0, −0.7, −0.5) rated 4,000 papers each (randomly selected), and the other eight raters rated 1,470 papers each (randomly selected).

Rater Distribution 3 (D3)

Three harsh raters (0.5, 0.7, 1.0) rated 4,000 papers each (randomly selected), and the other eight raters rated 1,470 papers each (randomly selected).

Rater Distribution 4 (D4)

Three moderate raters (−0.1, 0, 0.1) rated 4,000 papers each (randomly selected), and the other eight raters rated 1,470 papers each (randomly selected).

Evaluation Criteria

For each response data for all the rater distribution and replications, two sets of responses were created: one set contains both MC and CR items, and the other set contains CR items. Markov chain Monte Carlo estimates based on Yao's rater model for both sets were conducted.

It is expected that the results based on the test that contains both MC and CR items are better than those with CR items; moreover, many state assessment data contain both MC and CR items. However, the Muraki's rater effect model can only handle data containing CR items.

Root mean square error (RMSE), absolute mean bias (ABSBIAS), bias, and test reliability were computed for all parameters and were used to examine the parameter recovery rates.

Let f_{true} be the true parameter and let f_j be the estimated parameter from sample j, then

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(f_j - f_{true})^2}, \tag{6}$$

where n is the number of replications. Here f can represent ability, rater, or item parameters.

$$\overline{f} = \frac{1}{n}\sum_{j=1}^{n} f_j, \tag{7}$$

where $\overline{f}$ is the final estimate.

The bias and absolute mean bias are defined as the following:

$$BIAS = \frac{1}{n}\sum_{j=1}^{n}(f_j - f_{true}), \tag{8}$$

$$ABSBIAS = \frac{1}{n}\sum_{j=1}^{n}|(f_j - f_{true})|. \tag{9}$$

Bias and absolute mean bias were used to assess the parameter recovery, and reliability was calculated as the square of the correlation between the estimates and the true values in the study. The squared differences between the true parameters and estimated parameters and squared bias were also calculated and plotted to show results of parameter recovery and comparison among different rater distributions.

Results

Using Markov chain Monte Carlo method following Yao's rater model, with 15,000 iterations and 3,000 as the burn-in item, rater and ability parameters were estimated for each

response data set. Large numbers of iterations were chosen to ensure the convergence of the Markov chain Monte Carlo chains, as it was not feasible to test convergence for all the conditions. However, some Markov chain Monte Carlo chains were examined by plotting a trace plot and convergence was achieved. The final item, rater, and ability estimates were obtained by averaging over the 10 replications. The priors for each of the parameters in the Markov chain Monte Carlo BMIRTII software run were specified in a later section. The means and standard deviations of scores of the 10 simulated data are presented in Table 2. Under each rater distribution pattern, the means and standard deviations are very close to each other across the 10 simulated data.

Table 2

Total Constructed-Response Score Means and Standard Deviations

	Distribution 1		Distribution 2		Distribution 3		Distribution 4	
Data	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	16.40	10.05	18.39	10.31	14.53	9.68	16.34	9.67
2	16.39	9.94	18.29	10.40	14.51	9.68	16.33	9.71
3	16.43	10.00	18.43	10.34	14.52	9.69	16.33	9.70
4	16.39	9.99	18.39	10.33	14.53	9.69	16.32	9.67
5	16.41	10.03	18.46	10.38	14.56	9.71	16.46	9.76
6	16.43	10.01	18.44	10.35	14.53	9.70	16.33	9.70
7	16.41	9.99	18.40	10.33	14.48	9.71	16.32	9.73
8	16.41	9.98	18.43	10.27	14.54	9.67	16.31	9.72
9	16.44	9.99	18.42	10.32	14.55	9.75	16.35	9.69
10	16.41	9.99	18.41	10.35	14.55	9.71	16.38	9.74

Note. N = 23,760.

Parameter Recovery With and Without Rater Effects Based on Yao's Rater Model

In Tables 3 and 4, the true and estimated CR item parameters and the true and estimated rater parameters are presented.

Table 3

Difficulty Parameter Recovery of the Simulated 15 CR Items Based on Yao's Rater Model

Item	Generating	Estimated (Distribution 1)	Estimated (Distribution 2)	Estimated (Distribution 3)	Estimated (Distribution 4)
1	0.72	0.72	0.71	0.72	0.72
2	2.05	2.04	2.04	2.04	2.04
3	3.11	3.10	3.10	3.12	3.11
4	2.82	2.81	2.81	2.81	2.82
5	1.15	1.14	1.13	1.14	1.15
6	2.23	2.23	2.22	2.22	2.23
7	-1.06	-1.08	-1.08	-1.08	-1.07
8	-0.80	-0.80	-0.82	-0.80	-0.80
9	-1.02	-1.04	-1.05	-1.05	-1.03
10	-0.15	-0.17	-0.17	-0.17	-0.16
11	-0.27	-0.28	-0.30	-0.30	-0.29
12	0.23	0.22	0.21	0.23	0.22
13	1.51	1.51	1.48	1.49	1.49
14	-1.68	-1.68	-1.69	-1.68	-1.69
15	-2.65	-2.64	-2.65	-2.65	-2.67

Table 4

Rater Parameter Recovery Based on Yao's Rater Model

Item	Generating	Est.(Distribution 1)	Est.(Distribution 2)	Est.(Distribution 3)	Est.(Distribution 4)
1	-1.00	-1.00	-1.01	-1.00	-1.00
2	-0.70	-0.70	-0.70	-0.70	-0.70
3	-0.50	-0.50	-0.49	-0.49	-0.50
4	-0.20	-0.20	-0.20	-0.20	-0.19
5	-0.10	-0.09	-0.11	-0.10	-0.10
6	0.00	0.00	0.00	0.00	0.00
7	0.10	0.10	0.11	0.10	0.11
8	0.20	0.20	0.19	0.20	0.21
9	0.50	0.50	0.50	0.50	0.50
10	0.70	0.70	0.70	0.70	0.71
11	1.00	1.00	1.00	1.00	1.00

Note. Est. = estimated.

Overall, rater and item estimates recover very well for models with raters for the four distributions for CR-only tests. However, among the four distributions, the squared differences between the true parameters and estimated parameters are the smallest for Rater Distribution 1, followed by Distribution 4, 3, and 2 (see Figure 1). In terms of rater parameter recovery, Figure 2 shows that the squared differences between the true rater parameters and estimated parameters are the smallest for Distribution 1, indicating that Distribution 1 recover the best, followed by Distribution 3, 2, and 4.

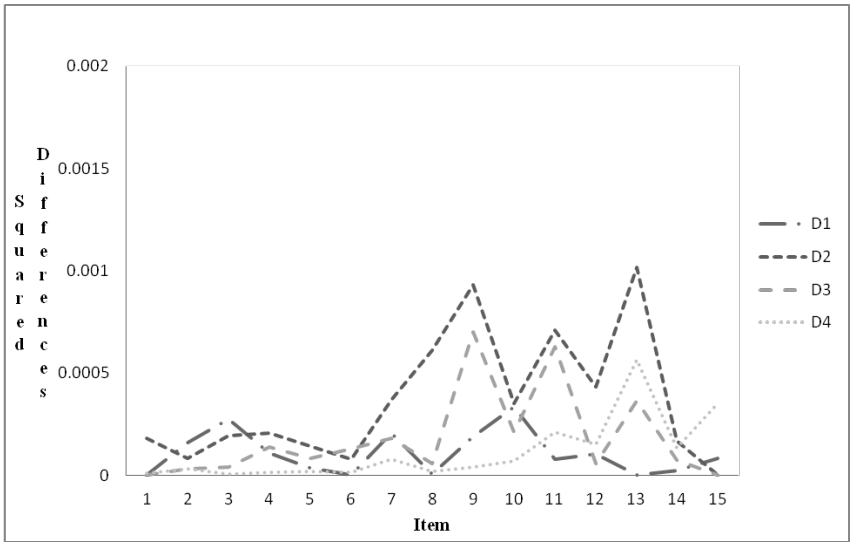

Figure 1. **Comparison of item difficulty parameter recovery across four rater distributions based on Yao's model.**

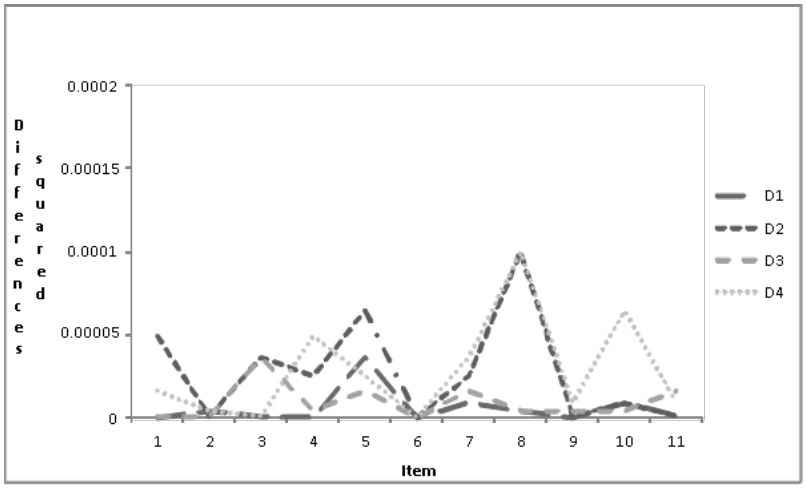

Figure 2. **Comparison of rater parameter recovery across four rater distributions based on Yao's model.**

In Table 5, results of the item difficulty parameter recovery comparison between with-rater parameter model and the with-no-rater parameter model are presented. When raters are included in the model, the estimated means and standard deviations are very close to the true means and standard deviations for all four distributions based on the 10 simulated data, and the correlations between the estimated and true scores are similarly close to 1.00. The item parameters do not recover well when using the with-no-rater model. The means become very small (e.g., drop from 0.41 to -0.60 for the Distribution 1 means) across the four distribution patterns. The means for the true and the estimates are quite different; however, the correlations between the estimates and the true are really high (> .99) for the four distribution patterns. This indicates that item difficulty values shifted away from the true values when they were estimated using an IRT model without rater parameters.

Table 5

Comparison of Item Difficulty With Raters and With No Raters, Based on Yao's Rater Model

CR only	Mean	SD	Mean	SD
Distribution	With no rater		With rater	
G1	0.41	1.72	0.41	1.72
1 EG1	-0.60	1.68	0.41	1.72
2 EG1	-0.55	1.71	0.40	1.72
3 EG1	-0.54	1.69	0.40	1.72
4 EG1	-0.52	1.71	0.41	1.72

Note. G1 is the generated *b*-parameter in Table 3; EG1 is the estimate for G1. $N = 15$.

Bias of Examinees' Ability Estimates for CR-Only and MC Plus CR Combined Tests Based on Yao's Rater Model

Comparisons were also conducted in terms of the squared bias of examinees' ability estimates between the four rater distributions at each ability level between –3.0 and 3.0 (see Figures 3 and 4). The results indicate that D4 has the largest bias at the lower end (< –2.0) and higher end (> 2.0) for CR-only and MC plus CR combined tests. However, D4 is very similar to the other three rater distributions at the middle range (–2.0, 2.0). D3 has slightly larger bias than D1 and D2 at the lower end for the CR-only tests. For MC plus CR combined tests, larger bias is

observed outside the ability range of (–2.0, 2.0); the bias is very small for the ability range of (-2, 2). D4 has a larger bias than the other three rater distributions. CR-only tests yield larger bias than MC plus CR combined tests.

Generally speaking, the examinee ability estimates with the no-rater model are very poor for CR-only tests. However, when MC plus CR items are combined, the impact of using the with-rater model or with-no-rater model is less; ABSBIAS, RMSE, and BIAS values do decrease to some extent, but not so much when they are compared to the values from CR-only tests.

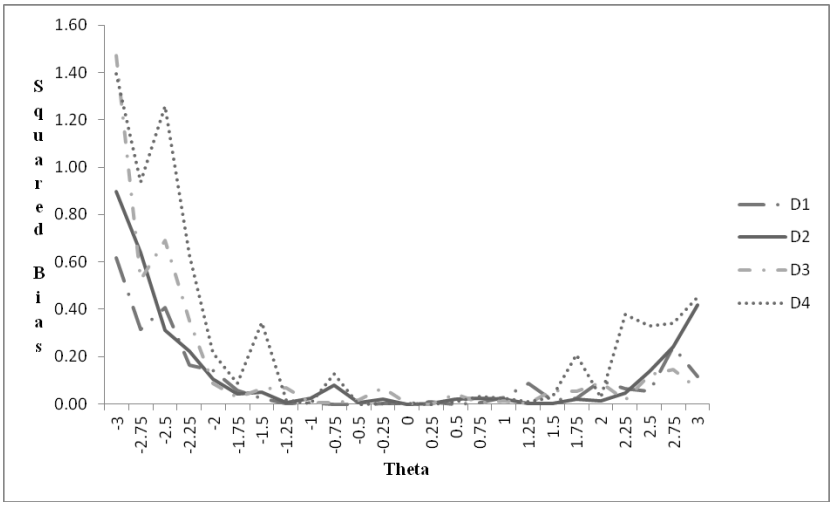

Figure 3. **Comparison of four rater distributions of CR-only test: 1PL model with rater based on Yao's model.**

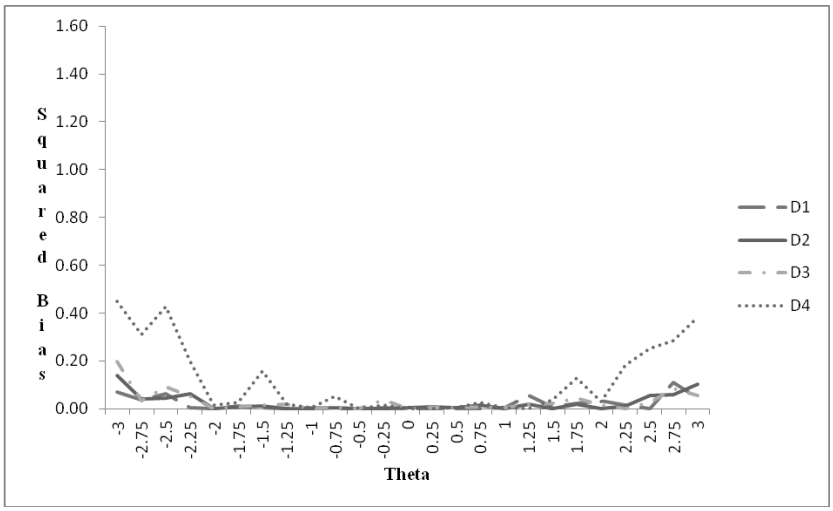

Figure 4. **Comparison of four rater distributions of MC plus CR tests: 1PL model with rater based on Yao's model.**

Comparison of Mean Estimates of Examinees' Ability Across the Four Rater Distributions Based on Yao's Model

To examine the ability estimate difference between the with-rater model and the with-no-rater model, Markov chain Monte Carlo 2PL parameter estimates for the 10 data sets were obtained. Table 6 shows absolute bias, RMSE, and bias values for the ability parameters, with-rater, and without-rater parameters for data sets with CR-only and MC plus CR combined tests.

Table 6 shows that absolute bias values are .25 to .76 without rater effects in the model but only .19 to .35 when rater effects are included. RMSE values are .14 to .35 without rater effects in the model but only .08 to .16 when rater effects are included. Reliability values range from .79 to .94 when rater effects were included; however, they range from .01 to .89 without rater effects in the model.

When rater effects are included in the model, in terms of the differences between the four rater distributions, Table 6 shows that the absolute bias and RMSE values for D4 are larger than those for D1, D2, and D3 for both CR-only and MC plus CR combined tests. For bias, D2 has the smallest values as compared with the other three distributions for both CR-only and MC plus CR combined tests.

When rater effects are not included in the model, in terms of the differences of the four rater distributions the results show that D1, D2, and D3 produce better ability estimates with smaller absolute bias, RMSE, and bias values than those of D4.

Comparison of Test Reliability Across the Four Rater Distributions

Reliability was also obtained and compared between the four rater distributions (see Table 6). We found that the reliability values for both MC plus CR items combined tests were higher than those with CR-only tests for the four rater distributions. In terms of with-rater model, for Rater Distributions 1, 2, and 3, the reliability values for CR-only tests are in the range between 0.86 and 0.87, which is much higher than that of Rater Distribution 4 (0.79). Similarly, for MC plus CR combined tests, the reliability values for Rater Distributions 1, 2, and 3 are around 0.94, which is much higher than that of Rater Distribution 4 (0.84). In terms of the with-no-rater model, the reliability values for the CR-only tests drop significantly, and they range from 0.08 to 0.15 for the four rater distributions. For MC plus CR combined tests, the reliability values drop from 0.94 to 0.89 for Rater Distributions 1, 2, and 3, and drop from 0.84 to 0.71 for Distribution 4.

Table 6

Mean Estimates of Examinee Ability RMSE and Bias Across the Rater Distributions Based on Yao's Rater Model

	Distribution	Abs. bias	RMSE	Bias	Reliability
CR only	With rater				
	1	0.28	0.11	-0.0012	0.87
	2	0.27	0.10	0.0004	0.88
	3	0.29	0.11	-0.0015	0.86
	4	0.35	0.16	-0.0010	0.79
MC+CR	With rater				
	1	0.20	0.08	-0.0014	0.94
	2	0.19	0.07	-0.0009	0.94
	3	0.20	0.08	-0.0015	0.94
	4	0.28	0.14	-0.0015	0.84
CR only	Without rater				
	1	0.74	0.34	-0.0022	0.08
	2	0.70	0.33	0.0022	0.15
	3	0.75	0.34	-0.0022	0.06
	4	0.76	0.35	-0.0021	0.04
MC+CR	Without rater				
	1	0.26	0.14	-0.0023	0.89
	2	0.25	0.14	-0.0024	0.89
	3	0.26	0.14	-0.0023	0.89
	4	0.42	0.26	-0.0026	0.71

Note. CR = constructed response; MC = multiple choice; RMSE = root mean squared error; Abs. = absolute.

Generally speaking, the reliability with the no-rater model is really poor for CR-only tests. However, when MC plus CR items are combined, the impact of using the with-rater model or the with-no-rater model is less; reliability values do decrease to some extent, but not so much when they are compared to the values from CR-only tests.

Model Comparison of Muraki's and Yao's Rater Model

Muraki's rater effect model. We used PARSCALE to run Muraki's rater effect model for CR-only tests. We selected normal on equally spaced points as the prior distribution. The method employed in the calibration phase of PARSCALE is that of random-effects estimation through marginal maximum likelihood. The random-effects solution employs the EM method of solving the marginal likelihood equations. We set the number of EM cycles to be 100, and the number of quadrature points as 30 in the EM and Newton estimation. The scale parameter was set at 1.0 in the PARSCALE run.

Yao's rater model. Yao's rater model run was based on the Markov chain Monte Carlo estimation method. The following prior values were set for the Markov chain Monte Carlo estimation:

1. Number of iterations: 15,000
2. Number of burn-in samples: 3,000
3. Prior for population: N (0, 1)
4. Prior for difficulty parameter b: N (0, 1.5)
5. Prior for rater parameter: N (0, 0.6)
6. Prior for guessing parameter beta: (100, 400)
7. The scale parameter: 1.0

We compared rater parameter recovery and examinees' ability estimates using Data 1 only. There are significant differences in terms of the item parameter estimates from the two models. The item parameters from Yao's rater model have much higher correlations with the true parameters than that of Muraki's rater effect model (see Table 7). The item parameters from Muraki's rater effect model have correlations ranging from .84 to .87, with the true parameters. In Figure 5, the ability estimates from the two models are plotted and have a strong positive linear relationship. In Figure 6, the differences of the ability estimates from the two models are plotted, and they are within the range of -1.0 to 1.0.

Table 7

Intercorrelations of Estimated Parameter and True Parameters Across the Two Models

	Yao's rater model		Muraki's rater effect model	
CR only	Beta1	Beta2	Beta1	Beta2
Distribution1				
Beta1	1.00		0.86	
Beta2		1.00		0.85
Distribution2				
Beta1	1.00		0.84	
Beta2		1.00		0.84
Distribution3				
Beta1	1.00		0.87	
Beta2		1.00		0.84
Distribution4				
Beta1	1.00		0.86	
Beta2	-	1.00		0.84

Note. Beta1 is the item difficulty; Beta 2 is the item difficulty threshold.

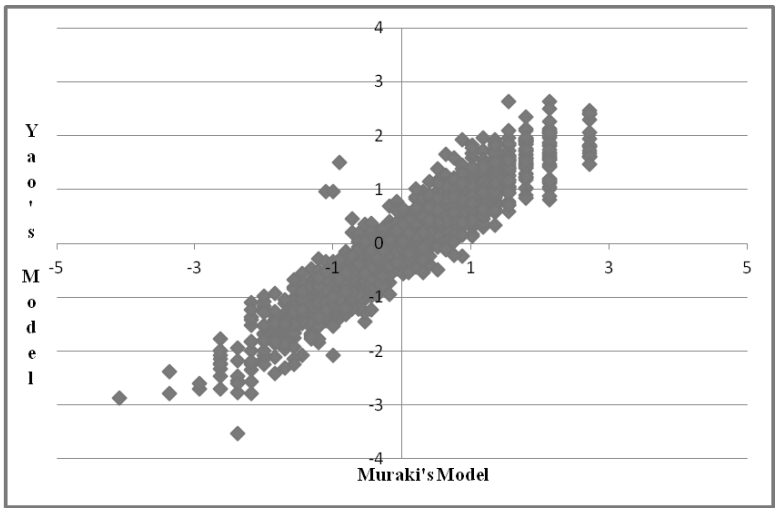

Figure 5. **Scatter plot of examinee ability estimates from Yao's and Muraki's rater effect models (Data 1: Sample *N* = 1,700).**

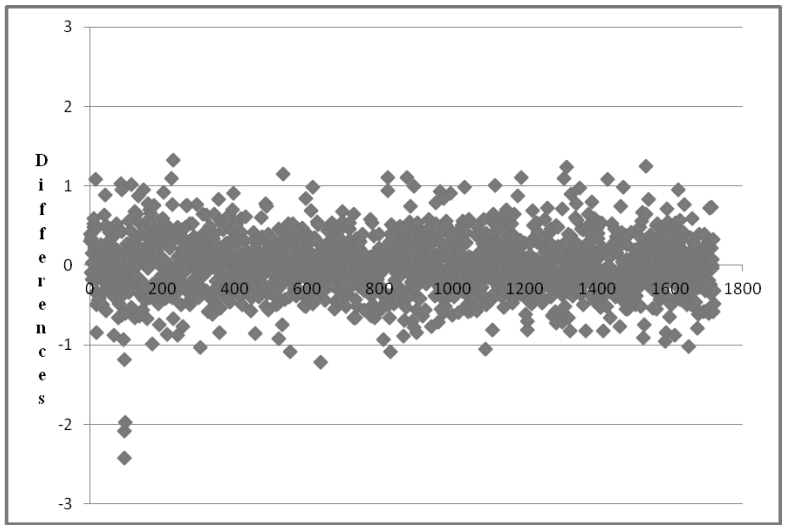

Figure 6. **Differences of examinee ability estimates between Yao's and Muraki's rater effect models (Data 1: Sample *N* = 1,700).**

In terms of rater parameter recovery, the rater parameter estimates from Muraki's rater effect model (see Table 8) are very close to the generated rater parameters, which is also similar to Yao's rater parameter estimates (see Table 4).

Table 8

Rater Parameter Recovery Based on Muraki's Rater Effect Model

Rater	Generating	Estimated (Distribution 1)	Estimated (Distribution 2)	Estimated (Distribution 3)	Estimated (Distribution 4)
	ρ_r	ρ_r	ρ_r	ρ_r	ρ_r
1	-1.00	-1.00	-0.97	-0.97	-0.95
2	-0.70	-0.66	-0.68	-0.69	-0.65
3	-0.50	-0.48	-0.48	-0.51	-0.52
4	-0.20	-0.18	-0.18	-0.21	-0.18
5	-0.10	-0.08	-0.09	-0.08	-0.11
6	0.00	-0.05	0.04	0.02	0.01
7	0.10	0.11	0.09	0.08	0.13
8	0.20	0.20	0.17	0.19	0.21
9	0.50	0.49	0.47	0.47	0.47
10	0.70	0.69	0.68	0.69	0.69
11	1.00	0.96	0.97	0.99	0.91

Conclusion and Discussion

Our study used simulated data to examine (a) the difference between Yao's IRT-based rater models with rater effect and with no rater effect, (b) the difference between the precision of the ability estimates for tests with CR only and with MC plus CR combined, and (c) the effect of four different rater distributions. We also compared Yao's rater model with Muraki's rater effect model in terms of ability estimates and rater parameters.

The recovery for the rater, item, and ability parameters for the rater models were examined using Yao's new IRT-based rater model. When rater parameters were included in the model, all parameters recover very well; however, when rater parameters were not included in the model, the ability estimates were extremely poor, especially for CR-only tests.

The results from the parameters' recovery are very similar across the four rater distributions. Distribution 1 recovered slightly better than the other three distributions. We also found that the effect of rater distributions and rater severity may increase the bias in the examinees' ability estimates to some extent. Our study was an initial attempt to examine those moderate raters whose scores tend to give middle-range scores. When moderate raters are assigned to read many more papers than other types of extreme raters, larger bias in the examinee ability estimates and lower test reliability can result. In our study, moderate raters seem to have more impact on examinees' ability estimate accuracy than the other rater effects, such as harshness or leniency, when raters are not assigned randomly during the rating process.

In this study, we also compared Yao's rater model with Muraki's rater effect model. We found that the ability estimates from the two models have a very strong linear relationship, although there are some differences that range from -1.0 to 1.0. The examinees' ability and item parameter estimates from Yao's rater model are better than those from Muraki's rater effect model. However, the rater parameter estimates from both models are very close to each other and are close to the generated values as well. The differences between Yao and Muraki's model comparison are typically the differences between Bayesian and maximum likelihood estimation. When comparing to maximum likelihood estimation, Bayesian estimation has more precision but larger bias, especially for small size (both length of the test and sample). When the sample size is large, the likelihood function dominates the posterior function; therefore, the effect of priors are small. The sample size (greater than 1,400) and the test length for this study are large, and thus the effect of prior should be small.

It seems that without random distributions of rater severity and ignoring rater effects can increase bias in the resulting report of examinees' final scores and reduce test reliability. Moreover, using IRT-based models without rater parameters being taken into account can substantially reduce test reliability and increase the bias in the resulting examinee ability estimates to some extent, especially when the test is composed of CR-only items. From our study, we found that the impact is less when the test is composed of both MC and CR items.

Results from this simulation study indicate that the distributions of rater severity can have some consequences for the accuracy of examinees' ability estimates and affect test reliability. Ignoring rater effects in choosing an analysis model can substantially reduce test reliability and increase bias in the resulting ability estimates of examinees, although the impact is less for the test that is composed of both MC and CR items. As Hombo et al. (2001) concluded, if an unlucky combination of extreme raters and examinees occurs in the scoring design, large bias in the examinee ability estimates can result.

Careful monitoring of the distributions of rater severity seems to be very critical in order to obtain high test reliability and minimize examinee ability estimation bias. Additionally, we recommend that decisions that are made based on the high-stakes assessment programs need to be re-evaluated if raters' performance is not modeled or checked. The results from the study offer a strong inducement for research to continue to develop and refine measurement models that correctly incorporate sources of measurement errors that may present in the data.

For future study, we will continue to conduct the simulation study using Yao's rater model with more parameters (e.g., discrimination parameters; more than one latent trait) to be estimated. Additionally, we can test the two rater models under various conditions of sample size (certainly number of examinees; possibly number of items/tasks), and possibly other conditions (random vs. nonrandom assignment). In this study, we used a rating design in which each rater was completely nested within a subsample of examinees, with each examinee assigned to one and only one rater. In order for rater effects models to be useful for placing parameter estimates onto the same scale under realistic conditions of small samples and nonrandom assignment, we can include some overlap across raters and examinees—some method for linking one subset of examinees and raters to other subsets of examinees and raters. Finally, we can use real operational data to test the robustness of Yao's rater model and Muraki's rater effect model.

References

Barr, M. A., & Raju, N. S. (2001). IRT-based assessments of rater effects in multiple source feedback instruments. *Organizational Research Methods, 6*, 15–43.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333–356.

de Jong, J., & Linacre J. M. (1993). Estimation methods, statistical independence, and global fit. *Rasch Measurement Transactions, 7*(2), 296–297.

Donoghue, J. R., McClellan, C. A., & Gladkova, L. (2006). *Using rater effects models in NAEP*. Unpublished manuscript.

Engelhard, G., Jr. (1996). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*, 19–33.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all examinees: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum.

Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics, 5*, 815–841.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (Research Report No. RR-01-05). Princeton, NJ: Educational Testing Service.

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

Linacre, J. M. (1991). FACETS [Computer program].

Longford, N. (1995). *Models for uncertainty in educational testing.* New York, NY: Springer-Verlag.

Muraki, E. (1993, April). *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.

Muraki, E., & Bock, R. D. (1999). PARSCALE 3 [Computer program]. Chicago, IL: Scientific Software International.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386–422.

Patz, R. J. (1997, April). *IRT and hierarchical approaches to rater variability in performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Patz, R. J., & Junker, B. W. (1999, April). *The hierarchical rater model for rated test items and its application to large-scale assessment data.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Patz, R. J., Junker, B. W., Johnson, M. J., & 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.

Sykes, R. C., Ito, K., & Wang, Z. (2008). Rater effects and the assignment of raters to items. *Educational Measurement: Issues and Practices, 27*(1), 44–45.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26,* 283–306.

Wilson, M., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*(1), 51–72.

Wolfe, E. W., & Myford, C. M. (1997, March). *Detecting order effects with a multi-faceted Rasch scale model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Yao, L. (2012, April). *Rater effect model*. Workshop presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*, 469–492.