Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 55

Rater Effects and Double Scoring in IRT Proficiency Estimation

Yoon Ah Song[†]
Brandon LeBeau
Won-Chan Lee

January 2021

[†] Yoon Ah Song is Psychometrician, Center for Applied Linguistics (email: epistteme84@hotmail.com). Brandon LeBeau is Assistant Professor, College of Education, University of Iowa (email: brandon-lebeau@uiowa.edu). Won-Chan Lee is Professor and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

All rights reserved

Contents

1 Introduction **1**

2 Item Response Theory for Rater Effects and Multiple Ratings **2**
2.1 Standard Polytomous IRT Models 2
2.2 IRT Models for Rater Effects and Multiple Ratings 4

3 Method **6**
3.1 Study 1: Single versus Double Ratings in the Standard IRT Model 6
3.2 Study 2: Double Ratings in IRT models 9

4 Results **9**
4.1 Single versus Double Ratings in the Standard IRT Model 9
4.2 Double Ratings in IRT models 14

5 Discussion **15**

6 References **17**

List of Tables

1 Frequencies of Not biased/ Biased Raters at Each of the Rater Effect Levels 19
2 Descriptive Statistics of Rater Parameters at Each of the Rater Effect Levels for Three-Score Category Items 20
3 Descriptive Statistics of Rater Parameters at Each of the Rater Effect Levels for Five-Score Category Items 20
4 Average of Bias, SE, and RMSE of Proficiency Estimates by Estimators, Number of Ratings 21
5 Effect Sizes (η^2) (%) of the Factors Explaining Variabilities in RMSEs of Proficiency Estimates 22
6 Average of Bias, SE, and RMSE by Sample Size, Estimators, and Presence of Rater Effects 23
7 Average of Bias, SE, and RMSE by Test Length, Estimators, and Presence of Rater Effects 23
8 Average of Bias, SE, and RMSE by Number of Score Category, Estimators, and Presence of Rater Effects 24
9 Average of Bias, SE, and RMSE by Number of Score Categories and IRT Models 24

List of Figures

1 Conditional Biases of Proficiency Parameter Estimates by Estimator and Item Score Treatment 25
2 Conditional SEs of Proficiency Parameter Estimates by Estimator and Item Score Treatment 26
3 Conditional RMSEs of Proficiency Parameter Estimates by Estimator and Item Score Treatment 27
4 Conditional Biases by IRT Model and Number of Score Category 28
5 Conditional SEs by IRT Model and Number of Score Category . 28
6 Conditional RMSEs by IRT Model and Number of Score Category 29

Abstract

Scoring constructed-response (CR) items often involves dealing with rater effects and multiple item scores. This paper presents the performance of IRT models when double ratings are used as item scores over single ratings when rater effects are present. Study 1 examined the influence of the number of ratings on the accuracy of proficiency estimation in the generalized partial credit model (GPCM), under varying levels of rater effects. Study 2 compared the accuracy of proficiency estimation of two IRT models (GPCM versus hierarchical rater model, HRM) for double ratings. The main findings were as follows: (1) rater effects substantially reduced the accuracy of IRT proficiency estimation; (2) double ratings relieved the negative impact of rater effects on proficiency estimation and improved the accuracy relative to single ratings; (3) IRT estimators showed different patterns in the conditional accuracy; (3) as more items and larger number of score categories were used, the accuracy of proficiency estimation improved; and (4) the HRM consistently showed better performance than the GPCM.

1 Introduction

In educational testing, types of test items can be divided into those for which the examinee must choose from a set of given options (selected-response, SR), and those for which examinees must construct answers. The focus of this current study is on constructed-response (CR) item types.

Although large-scale testing programs have made consistent efforts to incorporate CR items in their tests, including CR items has not been easy (Livingston, 2014). First, the number of CR items may be quite small, perhaps no more than five items, which directly affects the reliability of a test. Second, CR items are more vulnerable to breaches of test security because examinees taking CR items tend to put more time and effort into completing the task, which enables them to more easily remember the item(s). Most of all, concerns related to the rating process have been raised in the scoring procedure for CR items because scoring often relies on the subjectivity of human raters, leading to a source of variance referred to as rater effects.

As a broad category, rater effects lead to systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the examinee (p. 957) (Scullen, Mount, & Goff, 2000). Several types of rater effects have been identified in the literature, such as severity/leniency, centrality, and variability/inaccuracy (Engelhard, 1994; Wolfe, 2004; Wolfe & McVay, 2012). Severity/leniency is used to describe a pattern in which raters assign systematically lower/higher ratings than those that reflect the true rating of examinee performance. In contrast, centrality describes patterns in which raters tend to concentrate their ratings in the middle range of the scale. Rater variability/inaccuracy is used to describe random errors in rating behavior (Wolfe, 2004; Wolfe et al., 2012) that reduce consistency or reliability.

In practice, much effort has been made to control rating quality and to obtain high-quality ratings throughout the rating process (McClellan, 2010). By receiving extensive training, raters not only have content-specific knowledge, but also specific scoring procedures to evaluate responses to the CR items on a test. During the actual scoring process, several procedures for monitoring observed ratings are conducted: 1) back scoring, 2) validity scoring, 3) double scoring, and 4) trend scoring. In back scoring, assigned ratings are verified by experienced lead raters in rating teams. Representative answers for each scoring category are included among examinee responses and used to check rater performance in validity scoring. Trend scoring is required to maintain and apply the same scoring rubric over different occasions and is frequently used for equating tests with CR items.

Of these, in a double or multiple scoring system, all or part of the examinee responses are scored by at least two different raters (McClellan, 2010; Kim & Moses, 2013). As a result, it produces more than a single item score. The use of double ratings as item scores in item response theory models (IRT) is expected to increase the accuracy of examinee proficiency estimates; however, multiple ratings as an item score are only effective if there is evidence that it will have an impact on proficiency estimation and provide diagnostic information about

individual rater's rating performance (Patz, Junker, Johnson, & Mariano, 2002).

The primary purpose of this study is to investigate the effectiveness of using double ratings as item scores in IRT proficiency estimation when rater effects are present in CR items. Two simulation studies were designed and conducted. The first study compares the accuracy of proficiency estimation in the standard polytomous IRT model between single and double rating conditions. The second study compares the accuracy of proficiency estimation between two different IRT models for the double rating condition.

2 Item Response Theory for Rater Effects and Multiple Ratings

IRT allows for making consistent and valid inferences about items, test scores, or reliability measures (van der Linden, 2016). However, the IRT assumptions of local independence, dimensionality, and parameter invariance are needed for valid inference. In addition, choosing an appropriate functional form that includes specification of parameters and relationships among variables is important to avoid undesirable consequences given the purpose of the application of IRT models (Bolt, Deng, & Lee, 2014). Unfortunately, the practical consequence of this misfit is not necessarily aligned to the results of IRT goodness-of-fit indices (Bolt et al., 2014). When IRT models are used to fit multiple ratings, the impact of misspecification of rater effects or treating repeated ratings as independent might result in bias of parameter estimates (Boughton, Klinger, & Gierl, 2001; Park, 2011), underestimation of standard errors (Verhelst & Verstralen, 2001; Bock, Brennan, & Muraki, 2002; Patz et al., 2002; DeCarlo, Kim, & Johnson, 2011), or both.

IRT models for rater effects and multiple ratings can provide more information on the forms of rater effects at the individual rater level, while Generalizability theory (G-theory) provides test level information (Kim & Wilson, 2009). These models tend to fit CR items better by allowing the discrimination parameter to vary across items and can be more compatible with practical psychometric procedures such as scaling and equating (e.g. Kim, DeCarlo, & Lee, 2011).

2.1 Standard Polytomous IRT Models

The most frequently used polytomous IRT models for rating data sets are the graded response model (GRM, Samejima, 1969) and the generalized partial credit model (GPCM, Muraki, 1992). These models assume that only random errors are present in ratings without rater effects. So, the observed ratings are considered as direct responses from examinees.

In the generalized partial credit model (GPCM; Muraki, 1992), the item

score category function is expressed as:

$$p_{jk}(\theta_i) = \frac{\exp[\sum_{g=0}^{k} a_j(\theta_i - b_{jg}]}{\sum_{h=0}^{m_j-1} \exp[\sum_{g=0}^{h} a_j(\theta_i - b_{jg})]}, \quad (1)$$

where $p_{jk}(\theta_i)$ denotes the item score category function or response probability of examinee i ($i = 1, \ldots, N$) given latent ability level θ_i achieving score k ($k = 0, \ldots, m_{j-1}$) for item j ($j = 1, \ldots, T$). a_j is the discrimination parameter for item j and b_{jg} is the item score category difficulty parameter for gth category in item j. In the GPCM, item category difficulty parameters can be decomposed as $b_{jg} = b_j - d_{jg}$, where b_j indicates the item common difficulty parameter and d_{jg} indicates the relative difficulty parameter (or step parameter) specific to the gth category relative to other categories within item j, so they do not have to be ordered. In addition, $\sum_{g=0}^{0} a_j(\theta_i - b_{jg})$ is fixed to zero for identification.

There are no straightforward approaches for using standard IRT models for multiple ratings by multiple raters (Bock et al., 2002); however, several approaches can be applied. First, IRT models can be fitted as if multiple ratings were independent ratings from distinct items. In this case, the point estimates of proficiency parameters are expected to be unbiased; however, the standard errors of estimates are biased downward, which requires adjustment.

To correct this bias in standard errors, Bock et al. (2002) showed implementation of the correction factor in the likelihood function, which is defined as the ratio of the uncorrected measurement error variance (treating multiple ratings as independent) to the correct measurement error variance. However, this can only be applied to the raters nested within items design; otherwise the correction factor should be derived specific to the design used to collect data sets.

Another common approach is to use the linear combinations of multiple ratings such as through summation or averaging of multiple ratings as single item scores. In this approach, the local independence assumption will hold despite the use of multiple ratings in standard polytomous IRT models.

There are no differences statistically between summed and averaged item scores in terms of fitting polytomous IRT models, but they both require adjustment to the original item score categories. For example, for a three-score category CR item, item scores are 0, 1 or 2 when the item is single rated. If this item is rated by two raters and the ratings are summed, then the possible item scores are 0, 1, 2, 3, or 4. That is, item score categories are manipulated to be $2m_j - 1$ and two ratings per item and examinee are summed to be a single item score, where m_j is the original number of score categories for item j. For the averaged item scores, possible item scores are 0, 0.5, 1, 1.5, and 2. Since polytomous IRT models do not allow for decimals, these scores should be recoded into 0, 1, 2, 3, and 4, which creates the same item score categories in the summed item score method.

Summed or average scoring uses single item scores and does resolve the violation of local independence among multiple ratings. However, it still has an impact of underestimating the standard errors of proficiency estimates due to

the larger number of score categories summed over item score categories in the information function.

In the summed or average scoring approach, concerns related to rater effects still remain. It is assumed that location effects (e.g., severity/leniency) are counterbalanced by randomization of rater assignments to ratings to obtain unbiased estimates (Bock et al., 2002; Verhelst & Verstralen, 2001). However, in practice, raters are not completely randomly assigned to fully remove rater effects of severity/leniency (Barnett, 2005) and still other types of rater effects such as rater variability can affect parameter estimation (Boughton et al., 2001; Park, 2011).

2.2 IRT Models for Rater Effects and Multiple Ratings

In IRT models for rater effects and multiple ratings, the observed rating is conceptualized as a function of the examinee, item, and rater. In these models, Y_{ijr} refers a polytomously scored rating variable for examinee i ($i = 1, \ldots, N$) to item j ($j = 1, \ldots, T$) with m_j number of item score categories and rated by rater r ($r = 1, \ldots, R$). p_{ijkr} is used to denote the rating probability of rater r assigning score k ($k = 0, \ldots, m_{j-1}$) in the response of examinee i to item j.

In standard IRT models, the local independence assumption implies that item scores are independent conditional on latent ability θ. For example, in single rated data sets, the product of probabilities of any examinee with ability θ to score 2 in item 1 ($p_{jk}(\theta) = p_{12}(\theta)$) and score 1 in item 2 ($p_{jk}(\theta) = p_{21}(\theta)$) by any raters is obtained by multiplying $p_{12}(\theta) \times p_{21}(\theta)$. However, this only holds when "the scores in the examinee's response record represent distinct items drawn from the content domain." (Bock et al., 2002). In double rated data sets, when both raters 1 and 2 assigned score 2 in item 1, and raters 3 and 4 assigned scores 1 in item 2, the probability of an examinee having score 2 in item 1 from raters 1 and 2, and score 1 in item 2 from raters 3 and 4 is not necessarily equal to the product of probabilities, $p_{12}(\theta) \times p_{12}(\theta) \times p_{21}(\theta) \times p_{21}(\theta)$.

If the local independence is assumed with multiple ratings in standard IRT models, theoretically the amount of information could be inflated, regardless of the number of items. Therefore, the important features of IRT models for rater effects and multiple ratings are not only to model the individual ratings to improve accuracy in parameter estimation, but also to resolve the issue of downward standard errors of measurement.

In this study, the hierarchical rater model (HRM; Patz et al., 2002) was used as one of the representative IRT models dealing with rater effects and multiple ratings. The HRM was introduced to better reflect rating process by modeling individual ratings with rater effects and to solve the issue of downward standard errors of measurement when treating multiple ratings as independent conditioned on ability. In the HRM, the relationship between observed scores and population parameters were established using G-theory and multilevel modeling strategies were adopted to specify the parameters and their relationships.

In G-theory, the true score of an examinee is viewed as the expected value over multiple ratings and items, which creates a hierarchy. Patz et al. (2002)

expressed this hierarchy using the normal true score model below,

$$\begin{aligned} \theta_i &\sim i.i.d.N(\mu, \sigma_\theta) \\ \eta_{ij} &\sim i.i.d.N(\theta_i, \sigma_\eta) \\ Y_{ijr} &\sim i.i.d.N(\eta_{ij}, \sigma_Y). \end{aligned} \tag{2}$$

Y_{ijr} is the rating score on the response of examinee i to item j rated by rater r, which is normally distributed with the mean η_{ij} and standard deviation of σ_Y. η_{ij} is the ideal rating (will be discussed later) reflecting the expected score of all observed ratings distributed normally with the mean of θ_i and standard deviation of σ_η. θ_i is the true ability parameter of examinee i. μ is the population mean of the latent ability distribution of θ_i with a standard deviation of σ_θ. This hierarchical structure is then assessed in a two-stage modeling process: 1) signal detection theory model is used to relate the observed response probability of Y_{ijr} to the ideal rating η_{ij} and 2) polytomous IRT model is used to determine the ideal rating η_{ij} conditioned on the latent ability θ_i.

In the HRM, the observed ratings for each of the raters are modeled by the degree of rater location and variability given the ideal rating. By reparametrizing the signal detection theory model, the HRM can be expressed as:

$$p_{ijkr} = P(Y_{ijr} = k|\eta_{ij} = \eta) \propto \exp\{-\frac{1}{2\psi_r^2}[k - (\eta + \phi_r)]^2\}, \tag{3}$$

where p_{ijkr} is defined as the probability of rater r assigning rating score k for examinee is response to item j conditioned on the ideal rating η_{ij} ($\eta = 0, \ldots, m_{j-1}$); and η_{ij} indicates the ideal rating for the response of examinee i to item j. Here, the ideal rating is the mean of the observed ratings as shown in the hierarchical structure model (3). The expected rating across all possible observed ratings Y_{ijr} would be η_{ij}, where rater effects are counterbalanced with each other. Mariano (2002) describes η_{ij} as the rating that would have been assigned by an expert rater without any rater effects.

ϕ_r indicates the rater location effect which can be viewed as a noise that affects the location of the original mean parameter of a normal distribution, η. When $\phi_r > 0.5$ ($k > \eta$), this implies a positive bias, in which a rater assigns higher scores rather than the ideal rating. When $\phi_r < -0.5$ ($k < \eta$), this indicates a negative bias, so a rater would assign the rating in lower categories rather than the ideal rating. Patz et al. (2002) explained that the role of ψ_r is analogous to the standard deviation in a normal distribution where its value near 0 indicates high consistency and larger values mean low consistency in the rater's rating behavior. If ψ_r gets large, the probability of rater r assigning score, $\eta + \phi_r$ decreases rapidly to zero. So, it is desirable for a rater to have a value near 0 to indicate low variability/high reliability.

In the higher level, η_{ij} is modeled with item and latent ability parameters using polytomous IRT models. For example, the GPCM or other polytomous IRT models can be used in this level and can differ in terms of the number of score categories of items. Using the GPCM defined in Equation (1), this can be rewritten as,

$$p_{j\eta}(\theta_i) = \frac{\exp[\sum_{g=0}^{\eta} a_j(\theta_i - b_{jg}]}{\sum_{h=0}^{m_j-1} \exp[\sum_{g=0}^{h} a_j(\theta_i - b_{jg})]}. \quad (4)$$

3 Method

3.1 Study 1: Single versus Double Ratings in the Standard IRT Model

A total of five factors were chosen for comparisons: 1) the number of ratings with three levels of item score treatment (single item score for single ratings, summed item score for double ratings, and separate item score for double ratings) and 2) four levels of rater effects (low location – low variability, high location – low variability, low location – high variability, and high location – high variability). We also included three additional factors that are considered important in parameter estimation of polytomous IRT models: 1) two levels of item score categories (3 and 5), 2) two levels of test length (5 and 10), and 3) three levels of sample size (500, 1,000, and 3,000).

By combining factors by levels, 144 simulation conditions with rater effects were created. In addition, baseline conditions without rater effects were added to compare and interpret results. Ratings in the baseline conditions were generated from the GPCM assuming without rater effects, while ratings in simulation conditions were generated using the HRM with rater effects. Each condition was replicated 100 times.

Three sets of ability parameters were randomly generated from $normal(0, 1)$ for sample sizes of 500, 1000, and 3000. Because the focus of this study is to investigate the impact of simulation factors on the accuracy of proficiency estimation, the ability parameters were fixed across the simulation conditions.

Item parameters were randomly drawn from the statistical distributions but manipulated to avoid extreme values in consideration of the fact that items with extreme values are eliminated from item pools in practice. Item discrimination parameters were randomly selected from $uniform(0.5, 2.5)$. For five-score category items, four item difficulty parameters were randomly drawn from $uniform(-2.5, 2.5)$ for each item, and these item difficulty parameters were ordered within each item. For three-score category items, two difficulty parameters were drawn from $uniform(-2.5, 2.5)$ and ordered.

Three different approaches were applied to determine item scores depending on the number of ratings. First, single ratings were directly used as item scores because only one score is available for each response to each item. For double ratings, summed ratings are frequently used as item scores. However, this involves more score categories than the original item, which may in turn affect accuracy. Therefore, a separate item score condition was added to examine the effect of double ratings without increasing the number of score categories. In this case, individual double ratings were separately used as the item scores and two proficiency estimates were obtained. Then, the average of these two values

was taken as the final proficiency estimate.

In this study, the level of rater effects indicates how many raters are identified as biased raters. For example, in low/high location levels, a test is rated by a group of raters for which fewer/more raters are identified as biased, either lenient or severe, in their rating behavior. Similarly, in low/high variability levels, the rater group has fewer/more raters classified as inaccurate. It is also assumed that a test is rated by a group of 36 raters with two independent rater effect parameters: location (leniency/severity) and variability (inaccuracy/inconsistency).

In order to determine the number of raters identified as biased at each level, two levels (low and high) were set for each rater effect in this study. The number of biased raters at low levels was set to approximately 20 percent for location and 5 percent for variability. The number of biased raters at high levels was set to approximately 25 percent for both rater effects. Then, assuming two rater effects as independent, the number of raters at each rater effect level was obtained by multiplying the total number of raters by the ratio of raters in the location level and the ratio of raters in the variability level. For example, for the low location – high variability condition, the number of raters who were biased in both location and variability was determined by multiplying the total number of raters (36) by the ratio of biased raters in the low location level (0.2) and the ratio of raters biased in the high variability level (0.25): $36 \times 0.2 \times 0.25$, which is approximately 2 raters. The number of biased/not biased raters at each level is shown in Table 1.

Rater parameters were randomly sampled from the uniform distribution for both location and variability parameters according to the predefined numbers. Patz et al. (2001) described that raters are identified as biased if the absolute value of location parameter ϕ_r is greater than 0.5, and less accurate if the variability parameter ψ_r exceeds 0. The rater variability parameters were generated from $uniform(0.1, 0.75)$ for three-score category items and $uniform(0.3, 1.25)$ for five-score category items. Considering that there might exist a greater variability in larger score category items, the lower limit was set to be larger for the five category items. In addition, variability parameters were classified as biased if the correlation between ideal rating patterns (ratings patterns generated without rater effects) and observed rating patterns (rating patterns generated given rater effect parameters) were between 0.70 and 0.80. To represent raters who are not biased, variability parameters were only used if the correlation between ideal rating patterns and observed rating patterns was greater than 0.80.

Assuming that raters are not biased by more than one standard deviation of the observed rating distribution, the upper/lower limit of location parameters were also set to be ± 0.75 for three-score category items and ± 1.25 for five-score category items. Therefore, for three-score category items rater location parameters were randomly generated from either $uniform(0.5, 0.75)$ to represent lenient, or $uniform(-0.75, -0.5)$ to represent severe raters. For five-score category items, parameters were generated from $uniform(0.5, 1.25)$ to represent lenient, or $uniform(-1.25, -0.5)$ for severe raters. Otherwise, location parameters were generated from $uniform(-0.5, 0.5)$ for both score categories.

All rater effect parameters were sampled from the ranges specified above. For example, in the low location and low variability condition, 29 out of 36 raters were not biased but 7 raters were biased in the location effect. Therefore, for three-score category items, 29 raters out of 36 raters were assigned location values (ϕ_r) ranging $(-0.5, 0.5)$, and the remaining 7 raters were assigned values either from $(0.5, 0.75)$ or $(-0.75, -0.5)$.

Next, 36 location parameters were randomly paired with rater variability parameter values. Using rater effect parameters, ideal ratings and observed ratings were generated and the correlations were computed between ideal and observed ratings for each rater parameter pair. For example, when 2 of the 36 raters were classified as inaccurate, 2 rater effect parameter pairs were selected if the correlation was between 0.70 and 0.80. For the remaining 34 raters, rater effect parameters were used only if correlations were greater than 0.80. The same selection process was repeated until appropriate rater effect parameter pairs were found and applied to all four levels of rater effects at each item category. Descriptive statistics of rater effect parameters used for the simulation are summarized in Table 2 and Table 3.

To evaluate the accuracy of proficiency estimation, for each of the proficiency estimates $\hat{\theta}$, bias, standard error (SE), and root mean squared error (RMSE) were computed using the following formulas:

$$\text{bias}_{\hat{\theta}} = \frac{\sum_{I=1}^{100}(\hat{\theta}_I - \theta)}{100}, \tag{5}$$

$$\text{SE}_{\hat{\theta}} = \sqrt{\frac{\sum_{I=1}^{100}(\hat{\theta}_I - \bar{\hat{\theta}})^2}{100}}, \tag{6}$$

and

$$\text{RMSE}_{\hat{\theta}} = \sqrt{\frac{\sum_{I=1}^{100}(\hat{\theta}_I - \theta)^2}{100}}, \tag{7}$$

where θ is the true ability parameter and $\hat{\theta}_I$ is its estimate at iteration I using any of the four estimators: expected a priori (EAP or $\hat{\theta}_{EAP}$), summed score EAP (summed EAP or $\hat{\theta}_{sEAP}$), maximum likelihood (ML or $\hat{\theta}_{ML}$) or test characteristic functions (TCF or $\hat{\theta}_{TCF}$) (for details about IRT proficiency estimators, see Kolen & Tong, 2010). $\bar{\hat{\theta}}$ is the average of the estimated values over 100 replications.

Lastly, Univariate ANOVA was used for significance testing for the factors explaining variabilities in RMSEs across simulation conditions because RMSEs can be considered as the overall accuracy measure of proficiency estimators as they encompass both bias and SE. Then, orthogonal contrasts were planned to study where the most differences occur between different levels of factors considered in the study. The effect size indicator η^2 was considered and interpreted as significant when it is greater than 0.1%, where η^2 is defined as between group sums of squares for the factor of interest divided by the total sums of squares.

3.2 Study 2: Double Ratings in IRT models

In Study 2, two factors were considered for comparisons: 1) two levels of polytomous IRT models and 2) two levels of item score categories. To avoid possible unreliable effects from a short test length and small sample sizes, a fixed sample size of 1,000 and the test length of 10 were selected for Study 2. Low levels of rater effects (low location – low variability) were chosen to show how well IRT models detect minor rater effects and differentiate from each other. Four simulation conditions were explored with ten replications for each condition (selected due to time constraints) to obtain parameter estimates using Bayesian estimation with the MCMC method.

Data sets from Study 1 were subsetted and reused for Study 2 if they met the following conditions: 1) double ratings, 2) number of item score categories (3 and 5), 3) test length (10), 4) sample size (1,000), and 5) rater effects (low location – low variability). Because Study 2 only uses ten replications, every tenth data set (e.g., 1, 11, 21,...,91) was selected and used for Study 2. The same parameters from Study 1 were adopted for Study 2.

To fit the GPCM and HRM using the Bayesian MCMC method, the following specifications of prior distributions were adopted. For the GPCM, $\theta_i \sim normal(0,1)$, $(i = 1,\ldots,N)$, $a_j \sim lognormal(1.2, 1.44)$, $(j = 1,\ldots,T)$, and $b_{jk} \sim normal(0, 6.25)$, $(k = 0,\ldots,m_{j-1})$ were used.

For the HRM, at the first level of observed scores, $\phi_r \sim normal(0, 10)$ and $log\psi_r \sim normal(0, 10)$, $r = (1,\ldots,R)$ were used. For the next level of the HRM, the same prior distributions were used as in the GPCM only model.

A total of 170,000 iterations was used with 2,000 burn-in and 15 thinning across the three chains, so that the posterior distribution from each chain could have 10,000 samples. Then, the mean of the posterior distribution was adopted as the parameter estimates in keeping with the first simulation study.

Study 2 used the same evaluation criteria and computation formulas as Study 1: bias, SE, and RMSE. The same logic and computation formulas can also be used for the HRM, because the HRM and GPCM share item and ability parameters. The primary difference is that the HRM estimates item and ability parameters based on the expected scores (ideal ratings) for all possible ratings, while the GPCM directly uses the observed rating scores.

4 Results

4.1 Single versus Double Ratings in the Standard IRT Model

Table 4 presents the overall averages of bias, SE, and RMSE of proficiency estimates by rater effects and number of ratings (i.e., item score treatment methods for double ratings) across estimators. As expected, the average biases of proficiency estimates under the rater effects conditions were close to those of the baseline conditions, due to the random assignment of raters to ratings, with a consistent direction noted across different conditions. However, it was

difficult to assess the absolute magnitude of influence of rater effects or the number of ratings on the bias because biases with different signs were averaged and canceled out with each other.

The rater effect condition demonstrated substantially larger SEs compared to baseline conditions across estimators, suggesting a reduction in the precision of proficiency estimates. In the rater effect conditions, both double scoring methods showed improvement in precision from the single rating condition. Rater effects for single ratings increased SEs by 0.3057 (65%) for the ML estimator and by 0.2984 (59%) for the TCF estimator relative to baseline. For the ML estimator, the separate item score reduced SEs by 0.1085 (23%) and the summed item score improved SEs by 0.1282 (16%). For the TCF estimator, gains in precision were 0.1235 (15%) for the separate item score method and 0.1269 (25%) for the summed item score method. Overall, the two double scoring methods demonstrated comparable efficiency for both the ML and TCF estimators. For the ML estimator, the summed item score method improved precision by 0.0197 (4%) which is greater than the separate items score method. For the TCF estimator, the difference between the two item score methods was less than 1%.

The EAP and summed EAP estimators were less affected by rater effects than the ML and TCF estimators. From the baseline to single rating condition, SEs increased by 0.0846 (24%) and 0.0723 (19%) for the EAP and summed EAP estimators, respectively. Gains from using double scoring methods were however smaller than in the ML and TCF estimators. For the EAP estimator, double scoring improved SEs by 0.0612 (17%) for the separate item score method and by 0.0266 (7%) for the summed item score method from the single rating condition. For the summed EAP estimator, double scoring improved SEs by 0.0591 (16%) for separate and 0.023 (6%) for the summed item score method. For these two Bayesian estimators, the separate item score method was more effective in improving precision than the summed item score method in terms of the overall average SEs.

Consistent with the previous evaluation of bias and SE, RMSEs were larger for the rater effect conditions relative to the baseline. The difference in RMSEs between the baseline and single rating conditions were 0.3099 (65%) for the ML and 0.3018 (59%) for the TCF, while the differences for the EAP and summed EAP estimators were 0.1467 (37%) and 0.1356 (32%), respectively. Relative to single rating conditions, the use of double ratings improved accuracy for the ML by 0.1070 (22%) for the separate item scores method and 0.1300 (27%) for the summed item scores method, for the TCF by 0.1234 (24%) for the separate item scores method and 0.1283 (25%) for the summed item scores method, for EAP by 0.0565 (14%) for the separate item scores method and 0.0528 (13%) for the summed item scores method, and for the summed EAP estimator by 0.0546 (13%) for the separate item scores method and 0.0504 (12%) for the summed item scores method.

Figures 1 through 3 demonstrate the conditional results for bias, SE, and RMSE of proficiency estimates by the number of ratings (item score treatment methods for double ratings) for each of the estimators. Figure 1 illustrates that the Bayesian EAP and summed EAP estimators had biased estimates,

which overestimated parameters at the lower ranges of the ability scale and underestimated parameters at the upper ranges of the ability scale. The ML and TCF estimators showed similar patterns, which remain unbiased across the range of the ability scale. While the overall average of biases provided little information on the impact of rater effects and the number of ratings due to cancellation issues, conditional biases showed that rater effects increase the biases in proficiency estimation for all estimators.

For the EAP and summed EAP estimators, rater effects substantially increased the conditional bias. The conditional biases were largest for the single rating condition, followed by the separate item scores method, summed item scores method, and baseline conditions across the range of the ability scale. The differences in biases between the baseline and rater effect conditions (single rating, separate and summed for double rating) tended to be larger at the upper and lower ends of the ability scale than in the middle. In particular, the difference in biases between single and baseline conditions at the upper and lower ends of the ability scale was as high as 0.6. Both double scoring methods had smaller biases than the single scoring condition; with differences between the summed item score method and baseline conditions method as large as 0.4. Of note, the separate item score method showed no difference from the single rating condition. An explanation for this may be that the proficiency estimates from the separate item scores were obtained from the average of two single unreliable proficiency estimates.

Regardless of the presence of rater effects, the ML and TCF estimators tended to consistently provide unbiased estimates across the range of the ability scale, with only small differences in biases across the baseline, single rating, and two double rating item score treatment conditions except at both ends of the ability scale. Although the baseline condition had the least biased estimates, the differences between rater effects and baseline conditions were negligible in the ability range between -2.5 and 2.5. Outside of the range, the differences in biases between single rating and the baseline conditions were about ± 0.3. Double scoring methods showed smaller biases than the single scoring methods; the difference between the summed and baseline conditions was about 0.1.

Thus, single scoring increased biases for the EAP and summed EAP throughout the ability range, and double scoring methods had an effect on reducing conditional biases. For the ML and TCF estimators, single scoring increased the biases at the lower/ upper ends of the ability scale, but the gains from double scoring were small relative to the Bayesian estimators.

In Figure 2, rater effects added variation in proficiency estimates and reduced precision across the range of the ability scale and estimators. The ML and TCF estimators were affected more by rater effects than the EAP and summed EAP estimators. The largest differences between the single and baseline conditions was about 0.1 for two Bayesian estimators, while the largest difference was about 0.3 for non-Bayesian estimators.

For the EAP and summed EAP estimators, conditional SEs were largest for the single rating condition, followed by the summed item scores method, separate item scores method, and baseline conditions in the middle range from -

1.5 to 1.5. However, outside of this middle range, the separate item score method had even smaller conditional SEs than the baseline condition. The separate item score method appeared to be more effective in reducing conditional SEs than the summed item score method, while the summed item score method was more effective in reducing the size of biases.

For the ML and TCF estimators, SEs were larger than those of the other two Bayesian estimators. For the EAP and summed EAP estimators, conditional SEs remained below 0.5 across the range of the ability scale regardless of the presence of rater effects and number of ratings, while conditional SEs for the ML and TCF estimators were larger than 0.5 for both the single and double rating conditions. For these two estimators, conditional SEs were largest for the single rating condition, followed by the separate item scores method, summed item scores method, and baseline conditions in the middle range of the ability scale. However, the SEs were largest for the single rating condition, followed by the summed item scores method, separate item scores method, and baseline conditions in the upper and lower range of the ability scale.

Figure 3 contains plots for conditional RMSEs of proficiency estimates by estimators and the number of ratings (item score). For four estimators, rater effects had negative impacts on the accuracy of proficiency estimation. The use of double scoring methods (summed and separate item scores) certainly improved accuracy: differences between double scoring and baseline conditions were not small throughout the range of the ability scale.

Consistent with the results from the conditional bias and SE, the EAP and summed EAP estimators had smaller RMSEs in the middle range of the ability scale from -2.0 to 2.0, while the ML and TCF estimators had smaller RMSEs at the lower and upper range of the ability scale than other two estimators. For the EAP and summed EAP estimators, the largest differences were about 0.6 between the single and baseline conditions and 0.4 between the summed item score and baseline conditions. The separate item score method had slightly smaller RMSEs in the middle range of the ability scale, but still showed similar RMSEs to the single rating condition and larger RMSEs than the summed score method at the lower/upper range of the ability scale.

Conditional RMSEs for the ML and TCF estimators were similar to the conditional SEs, which was likely because the source of variations in RMSEs is mainly from standard errors rather than biases. The largest differences between the single and baseline conditions were approximately 0.3 for the ML estimator and 0.28 for the TCF estimator. The largest difference in RMSEs between two double rating conditions (summed and separate) and baseline conditions were about 0.19 for the ML and 0.17 for the TCF estimator. While the summed score was the more effective double scoring method, both double scoring methods reduced RMSEs similarly for these estimators. For the ML estimator, the two double scoring methods were comparable. For the TCF estimator, the separate item score method was more effective than the summed item score method at the lower/upper range of the ability scale.

Table 5 shows the effect sizes of the simulation factors by estimators in RMSEs of parameter estimates. The number of ratings showed effect sizes ranging

from 0.14% to 4.32%. Most of the variability resulted from the differences between the baseline and single rating rater effect conditions rather than from the differences between single and double rating conditions. Only a small portion of the variability in RMSEs was explained by single versus double rating conditions for the ML estimator ($\eta^2 = 0.10\%$). Neither double rating item score treatment method showed differences in RMSEs ($\eta^2 \leq 0.00\%$).

Sample size showed small effect sizes in RMSEs only for the ML and TCF estimators ($\eta^2 \leq 0.10\%$). Sample sizes did not have any interaction effects on proficiency estimation with the number of ratings ($\eta^2 \leq 0.00\%$) or other study factors. As seen in Table 6, increasing sample size contributed to reducing the bias from 500 to 1,000; however, adding more samples from 1,000 to 3,000 did not decrease bias as much as increasing sample size from 500 to 1,000 regardless of the presence of rater effects. SE and RMSE also show little changes across the different sample sizes.

Test length demonstrated small to large effect sizes in RMSEs ($3.59\% \leq \eta^2 \leq 33.55\%$) across estimators. Test length also demonstrated small interaction effects with the number of score categories ($\eta^2 = 3.78\%$ for ML and $\eta^2 = 4.14\%$ for TCF), rater effects ($\eta^2 = 0.60\%$ for ML and $\eta^2 = 0.43\%$ TCF) and the number of ratings ($\eta^2 = 0.22\%$ for ML and $\eta^2 = 0.23\%$ for TCF). Differences in RMSEs between 5 and 10 item conditions in three-score category items were larger than in five-score category items for the ML and TCF estimators. For these two estimators, the accuracy increased more for the 5-item condition than the 10-item condition across the rater effect conditions. The largest increase occurred between the baseline and single ratings and the largest decrease occurred between single and double rating item scores in RMSEs.

As shown in Table 7, biases were slightly reduced as the test length increased from 5 to 10 across estimators. Test length was the most effective factor for improving precision (SEs) of proficiency estimation. The overall average SEs improved when test length increased from 5 to 10 items by 0.20 for the ML and TCF estimators and 0.09 for the EAP and summed EAP estimators. Gains in overall accuracy by increasing test lengths in the baseline conditions led to a 27% to 44% reduction in RMSEs, while gains were 25% to 37% in the rater effect condition.

The number of item score categories had small to medium effect sizes for the Bayesian estimators (7.13% for EAP and 2.26% for summed EAP) and large effect sizes for the non-Bayesian estimators (36.47% for ML and 39.71% for TCF). The number of score categories showed small interaction effects with rater effects ($\eta^2 = 0.69\%$ for ML and $\eta^2 = 0.64\%$ for TCF) and the number of ratings ($\eta^2 = 0.35\%$ for ML and $\eta^2 = 0.41\%$ for TCF). The differences in RMSEs between three- and five-score categories remained similar across levels of rater effects for the EAP and summed EAP estimators. For the ML and TCF estimators, RMSEs increased more rapidly from the baseline to the rater effect conditions. Similarly, for these estimators, the differences between two score categories were larger in the single rating condition than in the baseline or summed and separate item score conditions.

As seen in Table 8, five-score category items reduced overall mean biases

by 0.0005 for the EAP and summed EAP estimators compared to three-score category items, while overall average biases increased by 0.01 to 0.03 for the ML and TCF estimators in both baseline and rater effects conditions. For SEs, five-score category items provided more reliable proficiency estimates than three-score category items. In the baseline condition, increasing the number of score category reduced the SEs by 0.09 for Bayesian estimators and 0.2 for non-Bayesian estimators, on average. In the rater effect condition, an increase with score category from 3 to 5 reduced the SEs by 0.05 for Bayesian estimators and 0.3 for non-Bayesian estimators on average. In both conditions, the differences in RMSEs between two score categories ranged from 0.11 to 0.13 for the EAP and summed EAP estimators and 0.18 to 0.34 for the ML and TCF estimators. Overall, increasing the number of score categories tended to improve RMSEs.

Rater effects had small to medium effect sizes, 6.55% for EAP, 1.98% for summed EAP, 8.49% for the ML, and 6.99% for TCF estimators in RMSEs. Among rater effect levels, most of the variabilities occurred between baseline and low-location and low-variability conditions ($1.92\% \leq \eta^2 \leq 7.27\%$). For the ML and TCF estimators, rater effects showed small interaction effects with the test length ($0.43\% \leq \eta^2 \leq 0.60\%$) and number of score categories ($0.64\% \leq \eta^2 \leq 0.69\%$). For these two estimators, when the number of score category is three and test length is five, RMSEs increased more between baseline and low–low rater effect conditions, and between the other rater effect conditions and high-high rater effect condition.

4.2 Double Ratings in IRT models

Table 9 provides the average bias, SE, and RMSE results of fitting the GPCM and HRM models to double ratings under the condition of low-level rater effects. The overall average bias of proficiency estimates did not differ between the two models and number of score categories. The HRM reduced the average bias about 0.0013 (9.0%) for three-score category items and even increased by 0.0004 (3.4%) for five-score category items compared to the GPCM with the summed item score method. As shown in Figure 4, the HRM was more effective in reducing the conditional bias for three-score category items at the upper/lower range of the ability scale. However, for both score categories, the differences in average bias between the two models were minor across the ability scale.

As seen in Table 9, the HRM also improved the overall average SEs from the GPCM by 0.0310 (8.0%) for the three-score category items and 0.0220 (7.1%) for the five-score category items. In Figure 5, the HRM had slightly smaller conditional SEs throughout the ability scale. Still the GPCM showed similar levels of SEs to that of the HRM and differences in SEs between two models remained similar across both item score categories.

Overall, RMSEs of proficiency estimates from the GPCM decreased by 0.052 (11.5%) for three-score category items and 0.026 (7.6%) for five-score category items by fitting the HRM to the same double ratings. As shown in Figure 6, the HRM improved the accuracy for three-score category items more than five-score category items. In particular, the HRM contributed more to reduce the RMSEs

at the upper/lower range of the ability scale for three-score category items.

Consistent to the previous findings in overall and conditional averages of performance measures, the variabilities in performance measures attributable to the two models were small ($\eta^2 = 2.37\%$ for RMSE). The number of score category had medium to large effect sizes of 10.86% for RMSEs. There were small interaction effects between two IRT models and the number of score category because the differences between two IRT models in SE and RMSE were slightly larger for three-score category items than five-score category items as shown in Table 9.

5 Discussion

This study provides some implications to be considered when 1) scoring CR tests involves rater effects; 2) double/multiple ratings are available as item scores; and 3) IRT models are used to fit ratings:

- Impact of rater effects on IRT proficiency estimation
- Benefits of using double ratings as item scores over single ratings in the accuracy of IRT proficiency estimation
- Relative performance of four different IRT proficiency estimators
- Effects of sample size, test length, and number of score category
- Relative performance of two IRT models: HRM versus GPCM

First, the presence of rater effects in rating data sets reduced the accuracy of proficiency estimation in standard IRT models despite the random assignment of raters and high agreement rates.

Second, using single ratings as item scores substantially decreased the accuracy of proficiency estimates when rater effects are present in ratings. Using double ratings as item scores in CR items improved the accuracy of IRT proficiency estimation relative to the single rating condition. For double ratings, the summed item score method reduced both biases and standard errors, while the separate item score method was effective in reducing standard errors. Overall, both double scoring methods contributed to improving accuracy either by reducing standard errors, biases, or both. Given the results, the use of double scoring is recommended to improve the accuracy of IRT proficiency estimates compared to single scoring.

Third, IRT proficiency estimators yielded different patterns in conditional accuracy. Still, double scoring demonstrated improved accuracy over single ratings for any of the proficiency estimators. For the EAP and summed EAP estimators, rater effects significantly increased biases at the lower and upper range of ability scale. For the ML and TCF estimators, SEs increased across the range of ability scale. In RMSEs, for the Bayesian estimators, conditional results showed that the summed item score method was a more effective method

than the separate item score method, while both double scoring methods were effective to a similar extent for the ML and TCF estimators.

Fourth, longer test length and larger number of score categories improved accuracy in IRT proficiency estimation. Sample size was not as effective as test length or number of score categories when the sample size was greater than 1,000. The presence of rater effects in ratings was detrimental to the accuracy of proficiency estimation rather than the increase in levels of rater effects. The accuracy of proficiency estimation in CR item tests can significantly improve with a longer test length, larger number of score categories with double ratings.

Fifth, the HRM consistently provided more accurate proficiency estimates than the GPCM using the summed score method. The differences between the two models were not large due to the low level of rater effects. Applying the HRM can be more advantageous if the level of rater effects is substantially high in ratings. The HRM can provide item parameter estimates without the adjustment of original score categories. Also, the HRM can be a good alternative for resolving the issues of small samples for some of the score categories when the summed score method was used for double ratings.

There are several limitations that should be considered regarding the interpretation and generalization of the findings. First, this study only took the simulation approach. Real data analysis should be conducted to validate the simulation study design and generalize the results. These include setting rater parameters that are derived from the real data analysis with the information on rater memberships. Randomly generated parameters from the statistical distribution also restrict the generalization of the study results, which might be different from the true characteristics of parameters. In addition, this study did not reflect the real scoring practice in its simulation procedure, which might have significantly affected results of the study. For example, the procedures for addressing disagreement in ratings from multiple raters might produce different results.

Another limitation is that the results of this study can be interpreted in two ways. For Study 1, although the use of both double ratings as item scores certainly improved the accuracy, double scoring did not completely remove the negative impact of rater effects from proficiency parameter estimation. For Study 2, the HRM provided more accurate proficiency estimates; however, it is hard to conclude that the differences were significant enough and the HRM is advantageous to model multiple ratings with rater effects over the GPCM using summed item scores without knowing the level of rater effects in real practice. The HRM can be a better model since it provided more accurate proficiency estimates; however, the small differences between the two models may still indicate that the GPCM is also robust to rater effects.

Third, the focus of this simulation study is on constructed-response (CR) tests, for which rater effects are expected to be present in scores. Representative examples can include writing or speaking assessments or other performance type of assessments of which scores are assigned by raters judgments based on the scoring rubric. In practice, CR items are more often included in a mixed-format type of tests where most of the test items are selected response (SR) type of

items and a few items are from CR items. In this type of test, IRT parameter estimation could be more robust than CR item tests to rater effects and the violation of the local independence assumption if a sufficient number of SR items is included in a test. Thus, it should be recognized that depending on the proportion of CR items in the tests, the impact of double scoring and rater effects on proficiency estimation might differ.

Fourth, this study only considered two unidimensional IRT models: GPCM and HRM. There are possibilities that the results would change depending on the choice of unidimensional IRT models. Depending the characteristics of data sets, multidimensional measurement models may be a better fit. If a test is a mixed-format, it is more likely to have multidimensional latent traits. Therefore, it could be further investigated as to what extent multiple ratings are beneficial in more complex settings such as multidimensionality of latent traits and how raters rating behaviors interact with the multidimensional structure of latent traits.

Lastly, this study only evaluated the accuracy of proficiency estimation of two IRT models to examine the effect of double scoring under rater effects. Although double scoring produced more accurate proficiency estimates over single scoring, this does not necessarily indicate that double scoring could yield significantly better results in other statistics such as decision-making statistics, reliability, for both IRT and non-IRT approaches. Therefore, prior to determining the scoring procedure, depending on the purpose and final outcome of the study, the effect of double scoring can be assessed.

6 References

Barnett, S. G. (2005). *Relative Performance of Scoring Designs for the Assessment of Constructed Responses.* Doctoral dissertation, Measurement, Evaluation, and Research Methodology, The University of British Columbia, Vancouver, BC. Canada.

Bock, R., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*(4), 364-375.

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement, 51*(2), 141-162.

Boughton, K., Klinger, D., & Gierl, M. (2001, April). *Effects of random rater error on parameter recovery of the generalized partial credit model and graded response model.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, Seattle, WA.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29.

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.

Kim, S., & Moses, T. (2013). Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests. *International Journal of Testing, 13*, 314-328.

Kim, S. C, & Wilson, M. (2009). A comparative Analysis of the Ratings in the performance Assessment Using Generalizability Theory and The Many-Facet Rasch Model. *Journal of Applied Measurement, 10*(4). 408-423.

Kim, Y, K., DeCarlo, L. T., & Lee, W. (2011, April). *On Implications of a Hierarchical Rater/Signal Detection Model for Linking with Constructed Response Items.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, LA.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29*(3), 8-14.

Li, Y., & Baser, R. (2012). Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patientreported outcomes assessments. *Statistics in medicine, 31*(18), 2010-2026.

Livingston, S. A. (2014). *Equating Test Scores (Without IRT). (2nd ed.).* Princeton, NJ: Educational Testing Service.

Mariano, L. T. (2002). *Information Accumulation, Model Selection and Rater Behavior in Constructed Response Student Assessments.* Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

McClellan, C. A. (2010). Constructed-Response Scoring Doing It Right. *R & D Connections. 13.* Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-177.

Park, Y, S. (2011). *Rater Drift in Constructed Response Scoring via Latent Class Signal Detection Theory and Item Response Theory.* Unpublished doctoral dissertation, Columbia University.

Patz, R. J., Junker, B. W., Johnson, M. S., & 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*, 341384.

Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.* In Proceedings of the 3rd international workshop on distributed statistical computing (Vol. 124, No. 125.10).

Table 1: Frequencies of Not biased/ Biased Raters at Each of the Rater Effect Levels

Rater effect	Location		Variabililty		Total
	Not biased	Biased	Not biased	Biased	
Low-location/ low-variability	29 (80%)	7 (20%)	34 (94%)	2 (6%)	36 (100%)
High-location/ low-variability	26 (72%)	10 (28%)	34 (94%)	2 (6%)	36 (100%)
Low-location/ high-variability	29 (80%)	7 (20%)	27 (75%)	9 (25%)	36 (100%)
High-location/ high-variability	26 (72%)	10 (28%)	27 (75%)	9 (25%)	36 (100%)

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, 34 (4, Pt. 2).

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Su, Y. S., & Yajima, M. (2015). *R2jags: Using R to Run 'JAGS'. R package version 0.5-7*.
Retrieved from: https://CRAN.R-project.org/package=R2jags.

Verhelst, N. D., & Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. V. Duijn, and T. A. B. Snijders (Eds.), *Essays on Item Response Modeling* (pp. 89-108). New York: Springer-Verlag.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35-51.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*, 3137.

Table 2: Descriptive Statistics of Rater Parameters at Each of the Rater Effect Levels for Three-Score Category Items

Rater effects	Parameter	Mean	SD	Min	Max
Low location - low variability	Location	0.042	0.417	-0.709	0.718
	Variability	0.325	0.150	0.102	0.664
	Correlation	0.876	0.067	0.715	0.979
High location - low variability	Location	-0.032	0.400	-0.614	0.688
	Variability	0.238	0.142	0.015	0.587
	Correlation	0.899	0.071	0.756	0.990
Low location - high variability	Location	0.050	0.380	-0.638	0.711
	Variability	0.382	0.151	0.112	0.607
	Correlation	0.850	0.060	0.750	0.980
High location - high variability	Location	0.054	0.439	-0.638	0.698
	Variability	0.372	0.156	0.112	0.647
	Correlation	0.834	0.064	0.730	0.975

Note. Correlation: correlation between true ratings (baseline) and observed ratings (with rater effects)

Table 3: Descriptive Statistics of Rater Parameters at Each of the Rater Effect Levels for Five-Score Category Items

Rater effects	Parameter	Mean	SD	Min	Max
Low location - low variability	Location	-0.014	0.449	-0.966	1.195
	Variability	0.568	0.149	0.311	0.887
	Correlation	0.900	0.048	0.774	0.961
High location - low variability	Location	-0.054	0.551	-1.146	1.192
	Variability	0.585	0.180	0.304	0.909
	Correlation	0.889	0.059	0.759	0.982
Low location - high variability	Location	-0.056	0.488	-1.124	1.065
	Variability	0.680	0.216	0.304	1.105
	Correlation	0.858	0.076	0.701	0.960
High location - high variability	Location	0.085	0.588	-1.143	1.149
	Variability	0.644	0.193	0.305	1.023
	Correlation	0.864	0.064	0.736	0.951

Note. Correlation: correlation between true ratings (baseline) and observed ratings (with rater effects)

Table 4: Average of Bias, SE, and RMSE of Proficiency Estimates by Estimators, Number of Ratings

Estimator	Rater effect	Number of ratings (Item score)	Bias	SE	RMSE
EAP	Baseline	Single	-0.0175	0.3582	0.3961
	Rater effect	Single	-0.0176	0.4428	0.5428
		Double (Separate)	-0.0176	0.3816	0.4863
		Double (Summed)	-0.0175	0.4162	0.4900
ML	Baseline	Single	-0.0189	0.4723	0.4784
	Rater effect	Single	-0.0075	0.7780	0.7883
		Double (Separate)	-0.0067	0.6695	0.6813
		Double (Summed)	-0.0102	0.6498	0.6583
Summed EAP	Baseline	Single	-0.0175	0.3735	0.4187
	Rater effect	Single	-0.0176	0.4458	0.5543
		Double (Separate)	-0.0176	0.3867	0.4997
		Double (Summed)	-0.0176	0.4228	0.5039
TCF	Baseline	Single	-0.0188	0.5063	0.5122
	Rater effect	Single	-0.0098	0.8047	0.8140
		Double (Separate)	-0.0096	0.6812	0.6906
		Double (Summed)	-0.0116	0.6778	0.6857

Note. Baseline: rater effect are not present in ratings;
Rater effects: rater effects are present in ratings;
Separate: final proficiency estimates are determined by averaging two proficiency estimates obtained from each of the single ratings;
Summed: proficiency estimates are obtained by summing two single ratings.

Table 5: Effect Sizes (η^2) (%) of the Factors Explaining Variabilities in RMSEs of Proficiency Estimates

Factor	EAP	ML	Summed EAP	TCF
Sample size (N)	0.01	0.06	0.04	0.05
500 versus 1000, 3000	0.01	0.00	0.04	0.01
1000 versus 3000	0.00	0.05	0.00	0.04
Test lengths (T)	11.84	33.55	3.59	31.55
Number of score category (C)	7.13	36.47	2.26	39.71
Rater effects (R)	6.55	8.49	1.98	6.99
Baseline versus ll, hl, lh, hh	6.37	7.27	1.92	5.95
ll versus hl, lh, hh	0.12	0.57	0.04	0.50
hl versus lh, hh	0.04	0.31	0.01	0.26
lh verus hh	0.03	0.33	0.01	0.28
Number of Ratings (NR)	0.46	4.32	0.14	4.30
Baseline versus single, double	0.46	4.23	0.14	4.30
Single versus double	0.00	0.10	0.00	0.00
Double (separate versus summed)	0.00	0.00	0.00	0.00
N*T	0.00	0.00	0.00	0.00
N*C	0.00	0.00	0.00	0.00
N*R	0.00	0.00	0.00	0.00
N*NR	0.00	0.00	0.00	0.00
T*C	0.05	3.78	0.02	4.14
T*R	0.01	0.60	0.00	0.43
T*NR	0.00	0.22	0.00	0.23
C*R	0.01	0.69	0.00	0.64
C*NR	0.00	0.35	0.00	0.41
R*NR	0.00	0.04	0.00	0.04

Note. ll: low-location and low-variability; hl: high-location and low-variability; lh: low-location and high-variability; hh: high-location and high-variability

Table 6: Average of Bias, SE, and RMSE by Sample Size, Estimators, and Presence of Rater Effects

Estimator	Sample size	Baseline			Rater effects		
		Bias	SE	RMSE	Bias	SE	RMSE
EAP	500	0.0457	0.3562	0.3977	0.0457	0.4132	0.5109
	1000	0.0126	0.3629	0.3949	0.0126	0.4169	0.5033
	3000	-0.0381	0.3570	0.3962	-0.0381	0.4125	0.5066
ML	500	0.0397	0.4718	0.4783	0.0505	0.6954	0.7049
	1000	0.0089	0.4798	0.4861	0.0201	0.7105	0.7219
	3000	-0.0379	0.4698	0.4758	-0.0273	0.6959	0.7058
Summed EAP	500	0.0456	0.3707	0.4195	0.0457	0.4160	0.5223
	1000	0.0126	0.3776	0.4170	0.0126	0.4209	0.5155
	3000	-0.0380	0.3726	0.4192	-0.0382	0.4180	0.5200
TCF	500	0.0400	0.5042	0.5104	0.0480	0.7149	0.7233
	1000	0.0091	0.5146	0.5209	0.0175	0.7319	0.7418
	3000	-0.0378	0.5039	0.5096	-0.0293	0.7187	0.7270

Note. Baseline: rater effect are not present in ratings; Rater effects: rater effects are present in ratings

Table 7: Average of Bias, SE, and RMSE by Test Length, Estimators, and Presence of Rater Effects

Estimator	Test length	Baseline			Rater effects		
		Bias	SE	RMSE	Bias	SE	RMSE
EAP	5	-0.0176	0.4048	0.4589	-0.0176	0.4466	0.5770
	10	-0.0174	0.3117	0.3333	-0.0175	0.3805	0.4357
ML	5	-0.0239	0.5657	0.5722	-0.0107	0.8565	0.8682
	10	-0.0139	0.3788	0.3845	-0.0056	0.5417	0.5504
Summed EAP	5	-0.0176	0.4193	0.4844	-0.0176	0.4478	0.5890
	10	-0.0174	0.3277	0.3531	-0.0176	0.3891	0.4496
TCF	5	-0.0239	0.6079	0.6142	-0.0143	0.8803	0.8903
	10	-0.0136	0.4047	0.4103	-0.0064	0.5622	0.5699

Note. Baseline: rater effect are not present in ratings; Rater effects: rater effects are present in ratings

Table 8: Average of Bias, SE, and RMSE by Number of Score Category, Estimators, and Presence of Rater Effects

Estimator	Score category	Baseline			Rater effects		
		Bias	SE	RMSE	Bias	SE	RMSE
EAP	3	-0.0177	0.4000	0.4528	-0.0178	0.4406	0.5691
	5	-0.0173	0.3164	0.3394	-0.0173	0.3865	0.4436
ML	3	-0.0169	0.5658	0.5729	0.0085	0.8556	0.8686
	5	-0.0209	0.3787	0.3838	-0.0247	0.5426	0.5500
Summed EAP	3	-0.0177	0.4173	0.4817	-0.0178	0.4461	0.5887
	5	-0.0172	0.3298	0.3557	-0.0175	0.3908	0.4499
TCF	3	-0.0152	0.6124	0.6190	0.0044	0.8909	0.9013
	5	-0.0223	0.4002	0.4054	-0.0250	0.5516	0.5589

Note. Baseline: rater effect are not present in ratings; Rater effects: rater effects are present in ratings

Table 9: Average of Bias, SE, and RMSE by Number of Score Categories and IRT Models

Score category	Model	Bias	SE	RMSE
3	GPCM	0.0144	0.3860	0.4540
	HRM	0.0131	0.3550	0.4020
5	GPCM	0.0116	0.3120	0.3440
	HRM	0.0120	0.2900	0.3180

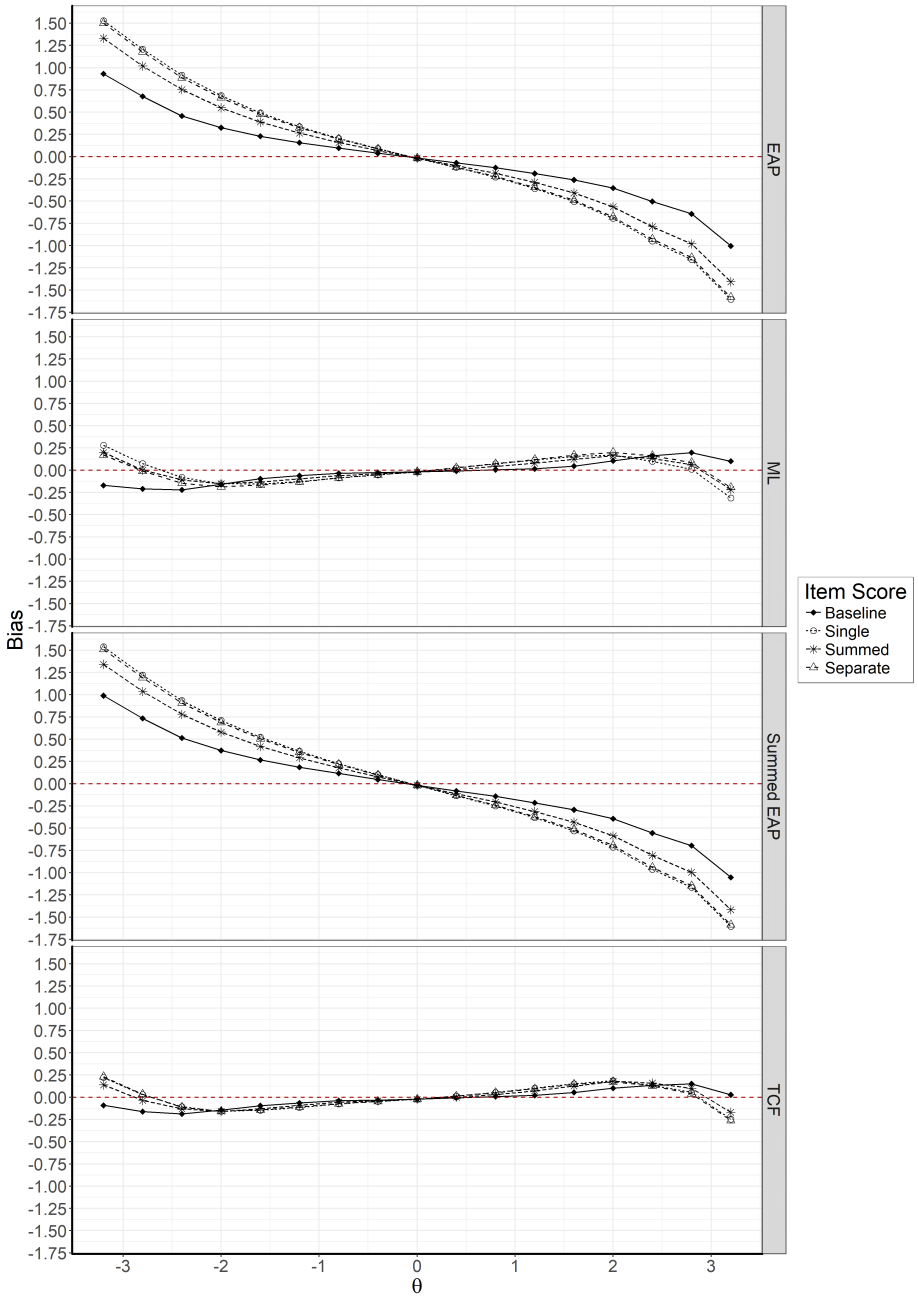

Figure 1: Conditional Biases of Proficiency Parameter Estimates by Estimator and Item Score Treatment

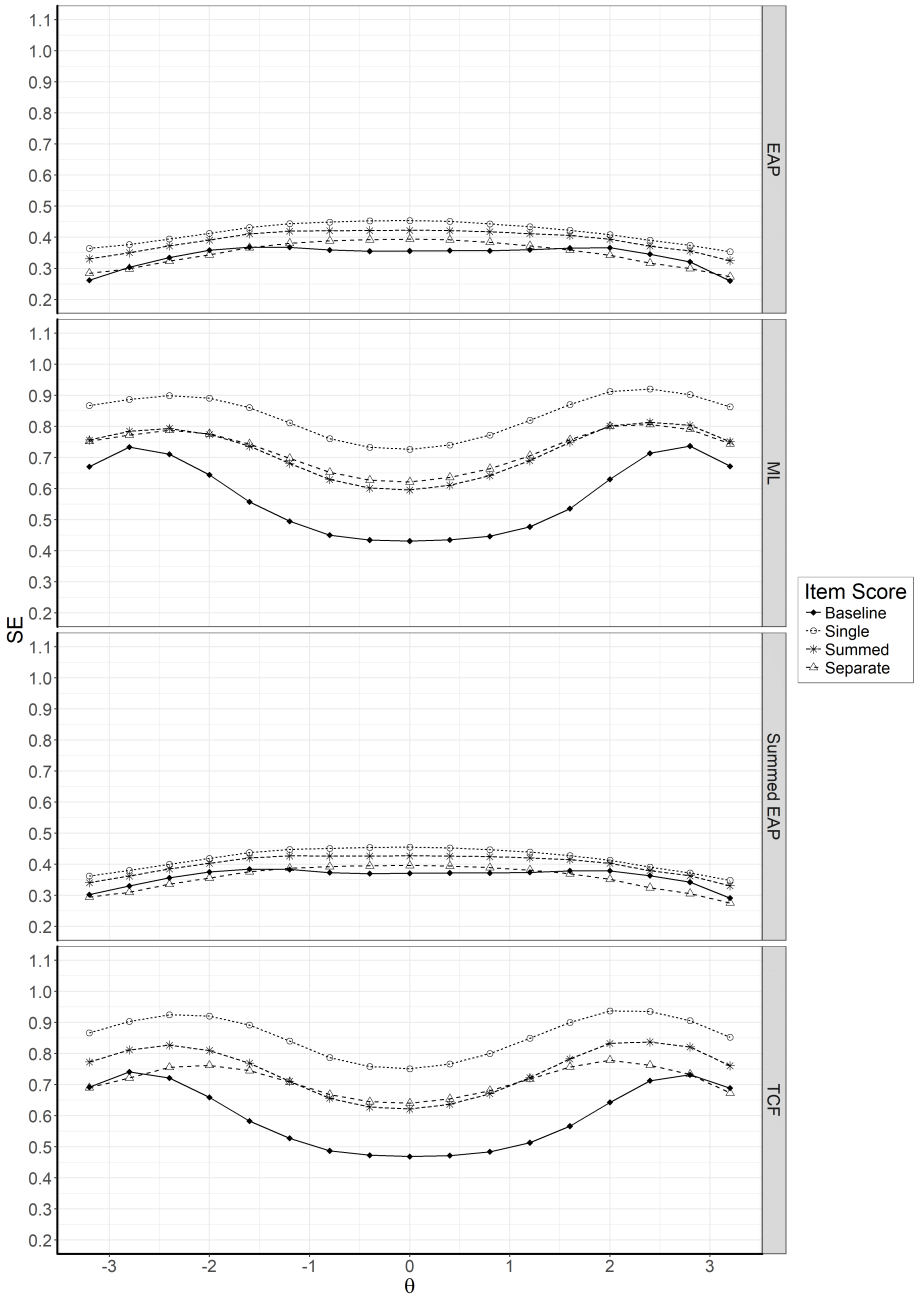

Figure 2: Conditional SEs of Proficiency Parameter Estimates by Estimator and Item Score Treatment

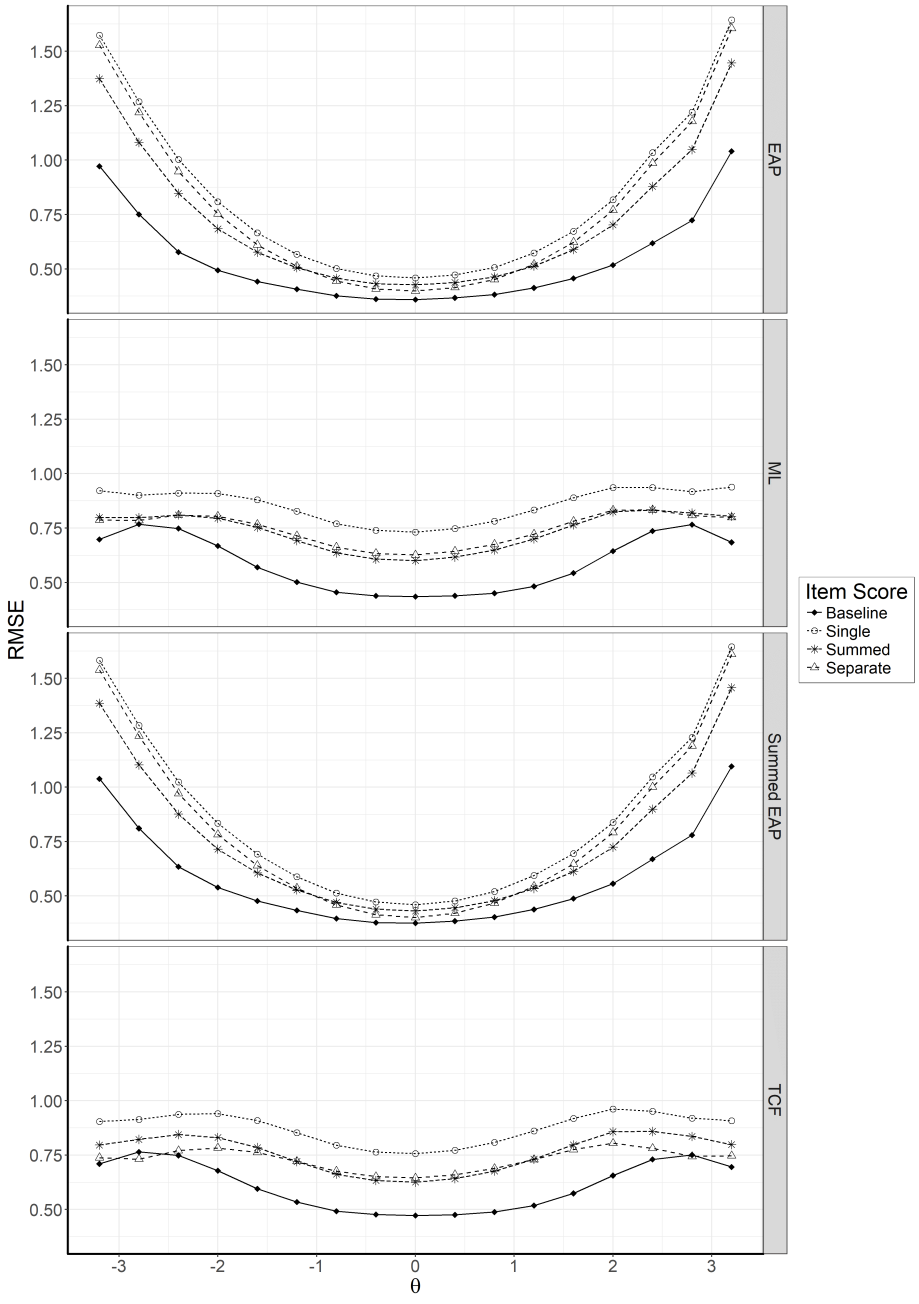

Figure 3: Conditional RMSEs of Proficiency Parameter Estimates by Estimator and Item Score Treatment

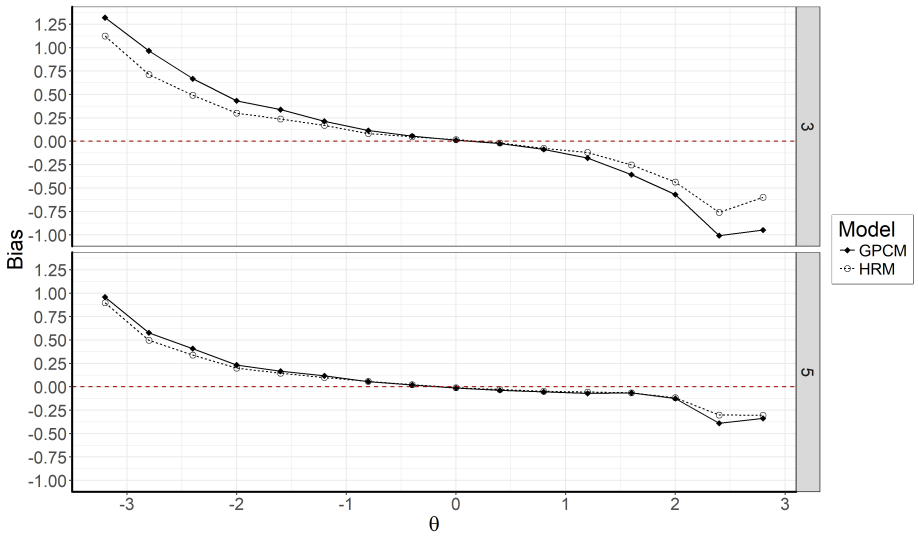

Figure 4: Conditional Biases by IRT Model and Number of Score Category

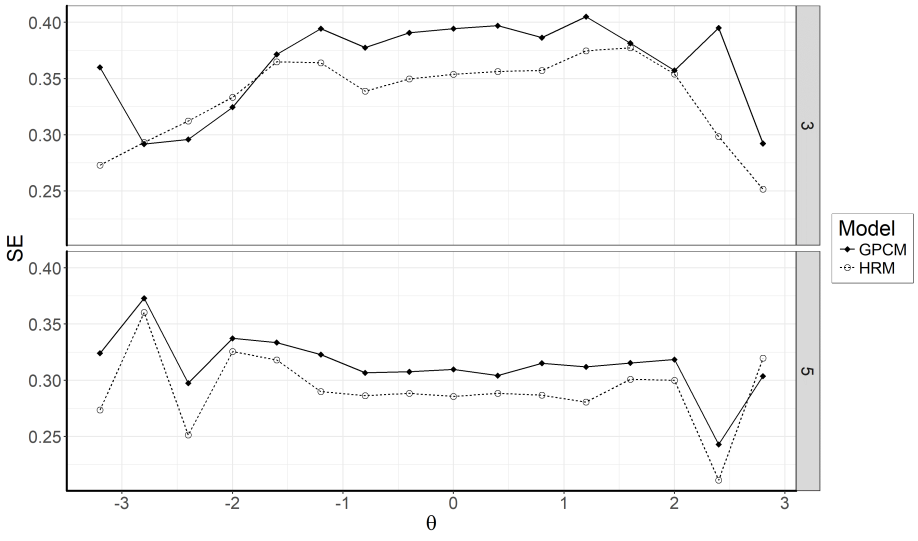

Figure 5: Conditional SEs by IRT Model and Number of Score Category

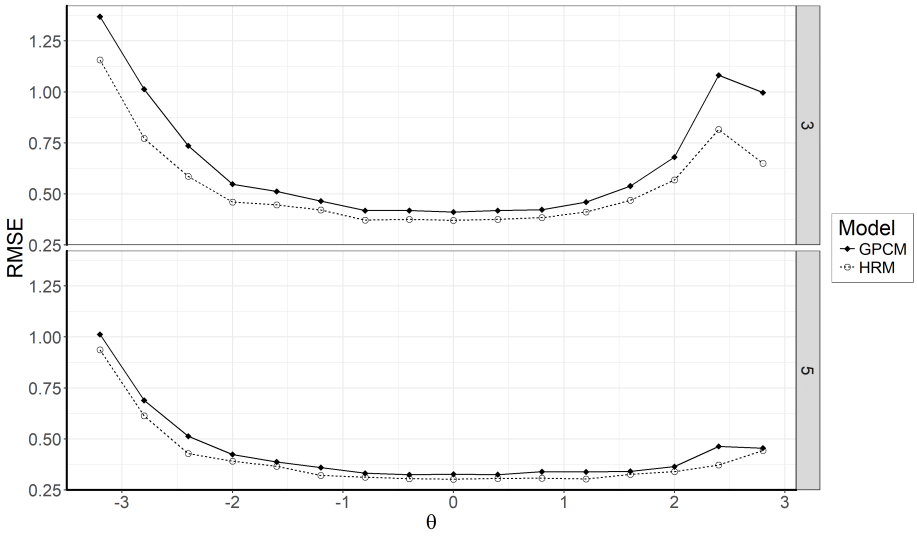

Figure 6: Conditional RMSEs by IRT Model and Number of Score Category