# Fitting Graded Response Models to Data with Non-Normal Latent Traits

**Tzu-Chun Kuo and Yanyan Sheng**

**Abstract** Fitting item response theory (IRT) models often relies on the assumption of a normal distribution for the person latent trait(s). Violating the assumption of normality may bias the estimates of IRT item and person parameters, especially when sample sizes are not large. In practice, the actual distribution for person parameters may not always be normal, and hence it is important to understand how IRT models perform under such situations. This study focuses on the performance of the multi-unidimensional graded response model using a Hasting-within-Gibbs procedure. The results of this study provide a general guideline for estimating the multi-unidimensional graded response model under the investigated conditions where the latent traits may not assume a normal distribution.

**Keywords** Polytomous item response theory • Multi-unidimensional graded response models • Hastings-within-Gibbs • Non-normal distributions

## 1 Introduction

Polytomous item response theory (IRT; Lord 1980) models are applicable for tests with items involving more than two response categories. Polytomous responses include nominal and ordinal responses. Ordinal polytomous responses, such as Likert scale items (Likert 1932), are broadly used in many fields, including education, psychology, and marketing. This study focuses on the graded response model (GRM; Samejima 1969), the most widely used IRT model for polytomous response data (e.g., Ferero and Maydeu-Olivares 2009; Rubio et al. 2007).

T.-C. Kuo (✉)
American Institute for Research, 1000 Thomas Jefferson Street, NorthWest, WA 20007, USA
e-mail: tkuo@air.org

Y. Sheng
Department of Counseling, Quantitative Methods, and Special Education, Southern Illinois University, Carbondale, IL 62901, USA
e-mail: ysheng@siu.edu



L.A. van der Ark et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 196, DOI 10.1007/978-3-319-56294-0_4

In many circumstances, multidimensional IRT (MIRT; Reckase 1997, 2009) models are adopted when distinct multiple traits are involved in producing the manifest responses for an item. A special case of the MIRT model applies to the situation where the instrument consists of several subscales with each measuring one latent trait, such as the Minnesota Multiphasic Personality Inventory (MMPI; Buchanan 1994). In the IRT literature, such a model is called the *multi-unidimensional* (Sheng and Wikle 2007) or the *simple structure* MIRT (McDonald 1999) model and is the major focus of the study.

The multi-unidimensional GRM applies to situations where a $K$-item instrument consists of $m$ subscales or dimensions, each containing $k_v$ polytomous response items that measure one latent dimension. With a probit link, the probability that the $i$th ($i = 1, 2, \ldots N$) person contains a Likert scale response with $c$ categories ($c = 1, 2, \ldots, C_j$) for the $j$th ($j = 1, 2, \ldots K$) item is defined as

$$\begin{aligned} P(Y_{vij} = c|\theta_{vi}, \alpha_{vj}, \boldsymbol{\delta}_j) &= \Phi(\alpha_{vj}\theta_{vi} - \delta_{j,c-1}) - \Phi(\alpha_{vj}\theta_{vi} - \delta_{j,c}) \\ &= \int_{\delta_{j,c-1}}^{\delta_{j,c}} \phi(z; \alpha_{vj}\theta_{vi})dz, \end{aligned} \tag{1}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF, respectively, $z$ is a standard normal variate, $\alpha_{vj}$ and $\theta_{vi}$ denote the item discrimination and the person's latent trait in the $v$th dimension ($v = 1, 2, \ldots, m$), and $\delta_{j,c}$ denotes the item threshold parameter for the $c$th response category of item $j$ (Samejima 1969), the latter of which satisfies

$$-\infty = \delta_{j,0} < \delta_{j,1} < \ldots < \delta_{j,C_j-1} < \delta_{j,C_j} = \infty. \tag{2}$$

From a theoretical perspective, latent trait distributions in the IRT literature are often assumed to be normal. Therefore, some common estimation methods, such as marginal maximum likelihood and Bayesian techniques, are developed assuming normal latent traits. However, in some psychological instruments, such as depression and anxiety tests, the population latent traits may follow a non-normal distribution. Research has shown that violating the assumption of normality may bias the estimates of IRT item and latent trait parameters (e.g., Sass et al. 2008; Reise and Revicki 2014). In the literature, studies have been conducted to investigate item and person parameter recovery in estimating unidimensional dichotomous (e.g., Kirisci et al. 2001; Sass et al. 2008) and unidimensional multi-group dichotomous (e.g., Santo et al. 2013) models, where the latent trait follows a non-normal distribution. However, little has been conducted to investigate parameter recovery in estimating multidimensional polytomous models in this regard.

In view of the above, this study focuses on investigating parameter recovery of estimating multi-unidimensional GRMs when latent traits are either normal or non-normal. Specifically, different distributions of person parameters are adopted, and the performances of estimating item and person parameters using Hastings-within-Gibbs (HwG; Kuo and Sheng 2015) are compared. The remainder of the paper is

outlined as follows. In Sect. 2, the HwG estimation is introduced. The simulation study is described and the results are discussed in Sect. 3. Finally, the conclusion for this study is summarized in Sect. 4.

## 2 Hastings-Within-Gibbs Estimation Procedure

For the past two decades, fully Bayesian has gained an increased popularity due to improved computational efficiency. There are two types of fundamental mechanisms among the Markov chain Monte Carlo (MCMC) algorithm: Gibbs sampling (Geman and Geman 1984) and Metropolis-Hastings (MH; Hastings 1970; Metropolis and Ulam 1949). Gibbs sampling is adopted in situations when the full conditional distribution of each parameter can be derived in closed form. If any of the full conditional distribution is not in an obtainable form, MH can be used via choosing a proposal or candidate distribution by the current value of the parameters. Then a proposal value is generated from the proposal distribution and accepted in the Markov chain with a certain amount of probability.

Hastings-within-Gibbs (HwG) is a form of the hybrid between Gibbs sampling and MH and has proved to be useful for complicated IRT models, such as GRMs. In the literature, Albert and Chib (1993) proposed a Gibbs sampler for the unidimensional GRM model. Cowels (1996) proposed a HwG procedure by using an MH step within the Gibbs sampler developed by Albert and Chib (1993) for sampling the threshold parameters to improve mixing and to accelerate convergence. Kuo and Sheng (2015) extended Cowles' approach to the more general multi-unidimensional GRM.

## 3 Simulation Study

To investigate parameter recovery of the HwG procedure in situations when latent traits are not normal, a Monte Carlo simulation study was carried out where tests with two subscales were considered so that the first half measured one latent trait ($\theta_1$) and the second half measured the other ($\theta_2$).

### *3.1 Simulated Data*

In the study, three factors were manipulated: sample size ($N$), test length ($K$), and intertrait correlation ($\rho$). The choice of $N$, $K$, and $\rho$ was based on previous studies with similar models. For example, when investigating multidimensional GRMs, Fu et al. (2010) adopted $N = 500, 1000$, $K = 10, 20, 30$, $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$ for dichotomous items and $N = 1000$, $K = 20$, $\rho = 0.2, 0.4, 0.6, 0.8$ for polytomous

items involving three categories. Working with dichotomous multi-unidimensional models, Sheng (2008) adopted $N = 1000$, $K = 18$, $\rho = 0.2, 0.5, 0.8$ in the simulation studies, while Sheng and Headrick (2012) adopted $N = 1000$, $K = 10$, $\rho = 0.2, 0.4, 0.6$. Wollack et al. (2002) conducted simulation studies with nominal response models, and they observed that parameter recovery was improved by increasing the test length from 10 to 30 items but that increasing the test length from 20 to 30 items did not produce a noticeable difference. Consequently, with our study, $N$ polytomous responses ($N = 500, 1000$) to $K$ items ($K = 20, 40$) were generated according to the multi-unidimensional GRM, where the population correlation between the two latent traits ($\rho$) was set to be 0.2, 0.5, or 0.8. Each item was set to be measured on a Likert scale with three categories so that two threshold parameters were estimated for each item. The item discrimination parameters $\boldsymbol{\alpha}_v$ were generated randomly from uniform distributions so that $\alpha_{vj} \sim U(0, 2)$. The threshold parameters $\delta_{j1}$ and $\delta_{j2}$ were sorted values based on those randomly generated from a standard normal distribution, i.e., $\delta_{j1} = \min(X_1, X_2)$ and $\delta_{j2} = \max(X_1, X_2)$, where $X_1, X_2 \sim N(0, 1)$.

The person parameters of the first dimension ($\boldsymbol{\theta}_1$) and the second dimension ($\boldsymbol{\theta}_2$) were generated based on the Method of Percentile (MOP; Koran et al. 2015) Power Method transformation. The MOP transformation was developed to generate multivariate distributions with specified values of median, interdecile ranges, left-right tail-weight ratios (a skewness function) and tail-weight factors (a kurtosis function) for each distribution, and the pairwise correlations.

To generate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ using the MOP transformation, $\boldsymbol{\theta}_1$ were generated from a standard normal distribution, and $\boldsymbol{\theta}_2$ were generated from one of the following four distributions: (1) skewness = 0, kurtosis = 0 (Dist. 1), (2) skewness = 0, kurtosis = 25 (Dist. 2), (3) skewness = 2, kurtosis = 7 (Dist. 3), and (4) skewness = 3, kurtosis = 21 (Dist. 4). The correlation between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ (i.e., the true intertrait correlation, $\rho$) was set to be 0.2, 0.5, or 0.8. Note that the skewness and kurtosis considered in each of the four distributions are conventional values and they can be transferred to left-right tail-weight ratios and tail-weight factors in order to implement the MOP transformation technique (see Koran et al. 2015).

Harwell et al. (1996) suggested that a minimum of 25 replications for Monte Carlo studies in IRT-based research is needed in order to obtain a better accuracy. Therefore, this study carried out 25 replications for each scenario, where root-mean-squared differences (RMSDs) and bias were used to evaluate the recovery of each item parameter. Let $\pi$ denote the true value of a parameter (e.g., $\alpha_{vj}$ or $\delta_{j,c}$) and $\hat{\pi}_r$ is the estimate in the $r$th replication ($r = 1, \ldots, R$). The RMSD is defined as

$$RMSD_\pi = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\pi}_r - \pi)^2}{R}}, \tag{3}$$

and the bias is defined as

$$bias_\pi = \frac{\sum_{r=1}^{R}(\hat{\pi}_r - \pi)}{R}. \tag{4}$$

The 10% trimmed means of these measures were calculated across items to provide summary statistics.

### 3.2 Results

Tables 1, 2, 3, and 4 display the results of the simulation study under the twelve test situations. The results indicated that the HwG procedure had an overall better estimation when $\boldsymbol{\theta}_2$ followed a normal distribution. The non-normality of $\boldsymbol{\theta}_2$ affected the accuracy of estimating $\boldsymbol{\alpha}_2$. Specifically, distributions 2–4 had overall larger RMSDs of $\boldsymbol{\alpha}_2$ than distribution 1 (normal). $\boldsymbol{\alpha}_1$ had similar RMSDs across these four distributions when $\rho = 0.2$ or 0.5. However, the non-normality of $\boldsymbol{\theta}_2$ had more influence on estimating $\boldsymbol{\alpha}_1$ when the two dimensions were highly correlated (i.e., $\rho = 0.8$). On the other hand, the performance of estimating $\boldsymbol{\delta}$ was affected more by skewness than kurtosis. Specifically, even though distribution 2 had the heaviest kurtosis, its RMSDs for estimating $\boldsymbol{\delta}$ were smaller than those from skewed distributions (i.e., distributions 3 and 4). The estimation of $\rho$ was sensitive to both skewness and kurtosis. Distributions 2–4 had larger RMSDs in estimating $\rho$ than distribution 1. A further comparison of its RMSDs under the four distributions indicated that they were similar when $\rho = 0.2$ but became more different when the actual correlation was higher (i.e., 0.5 or 0.8).

Posterior estimates for the person parameters ($\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$) were also obtained and correlated with their corresponding true values. Tables 1, 2, 3, and 4 summarize all the correlation results, where $r(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1)$ and $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ represent the correlations between the posterior estimates ($\hat{\boldsymbol{\theta}}$) and their corresponding true values ($\boldsymbol{\theta}$) for dimensions 1 and 2, respectively. The results indicate that $\boldsymbol{\theta}_1$ was estimated fairly well due to the satisfaction of normality assumption. On the other hand, the estimation of $\boldsymbol{\theta}_2$ was affected by kurtosis more than skewness, as distribution 2 had an overall lower $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ than distribution 3 (less kurtotic but more skewed). However, extreme skewed distributions (i.e., distribution 4) had an overall lower $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ than distributions 2 and 3. In addition, a comparison of $K = 40$ and $K = 20$ for the same sample size conditions (i.e., Table 2 vs. Table 1 and Table 4 vs. Table 3) indicates that the former had consistently larger $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ values than the latter. This suggests that the accuracy of estimating $\boldsymbol{\theta}_2$ improved with the increase in test length regardless of its distribution.

Further, it is found that an increase of sample size can improve the accuracy of estimating model parameters. For example, with the test length of $K = 20$, the RMSDs of estimating $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\rho$ when $N = 1000$ were in general smaller than those when $N = 500$, especially when the true intertrait correlation was higher. One shall note that when $\rho = 0.2$, larger sample sizes helped reduce the RMSDs of $\boldsymbol{\alpha}_2$ when $\boldsymbol{\theta}_2$ was non-normal. This is however not observed with $\rho = 0.5$ or 0.8. In terms of estimating $\boldsymbol{\theta}$, larger sample size tended to increase the accuracy of estimating $\boldsymbol{\theta}_1$. This pattern is only observed when estimating $\boldsymbol{\theta}_2$ in distributions 2 and 3 when $\rho < 0.8$.

**Table 1** Average RMSD and bias (italic values) for estimating $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\rho$ when $N = 500$, $K = 20$

| | True $\rho = 0.2$ | | | | True $\rho = 0.5$ | | | | True $\rho = 0.8$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{\theta}_2$ | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 |
| $\boldsymbol{\alpha}_1$ | 0.1248 | 0.1337 | 0.1264 | 0.1304 | 0.1242 | 0.1346 | 0.1248 | 0.1310 | 0.1191 | 0.1378 | 0.1395 | 0.1346 |
| | *0.0597* | *0.0708* | *0.0657* | *0.0707* | *0.0607* | *0.0739* | *0.0680* | *0.0724* | *0.0556* | *0.0873* | *0.0896* | *0.0863* |
| $\boldsymbol{\alpha}_2$ | 0.1088 | 0.1653 | 0.1659 | 0.1664 | 0.1177 | 0.1485 | 0.1468 | 0.1488 | 0.1261 | 0.1538 | 0.1527 | 0.1543 |
| | *0.0036* | *0.0061* | *0.0041* | *0.0048* | *0.0454* | *0.0344* | *0.0308* | *0.0332* | *0.0431* | *0.0589* | *0.0594* | *0.0601* |
| $\boldsymbol{\delta}_1$ | 0.1139 | 0.1217 | 0.1223 | 0.1296 | 0.1444 | 0.1483 | 0.1506 | 0.1558 | 0.1032 | 0.1002 | 0.1122 | 0.1138 |
| | *−0.0988* | *−0.0787* | *−0.0775* | *−0.0849* | *−0.0716* | *−0.0815* | *−0.0856* | *−0.0890* | *−0.0623* | *−0.0604* | *−0.0679* | *−0.0656* |
| $\boldsymbol{\delta}_2$ | 0.1124 | 0.1154 | 0.1193 | 0.1263 | 0.1256 | 0.1373 | 0.1436 | 0.1473 | 0.1041 | 0.1012 | 0.1026 | 0.1051 |
| | *−0.0543* | *−0.0587* | *−0.0569* | *−0.0656* | *−0.0574* | *−0.0620* | *−0.0651* | *−0.0706* | *−0.0296* | *−0.0352* | *−0.0391* | *−0.0398* |
| $\rho_{12}$ | 0.0026 | 0.0033 | 0.0030 | 0.0031 | 0.0029 | 0.0091 | 0.0079 | 0.0082 | 0.0005 | 0.0118 | 0.0116 | 0.0116 |
| | *0.0120* | *0.0265* | *0.0248* | *0.0256* | *0.0166* | *0.0794* | *0.0762* | *0.0772* | *0.0037* | *0.1066* | *0.1055* | *0.1059* |
| $r(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1)$ | 0.9160 | 0.9016 | 0.9090 | 0.9045 | 0.9100 | 0.9058 | 0.9141 | 0.9096 | 0.9394 | 0.9336 | 0.9336 | 0.9336 |
| $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ | 0.9170 | 0.7875 | 0.8420 | 0.7878 | 0.9148 | 0.7914 | 0.8436 | 0.7912 | 0.9290 | 0.8190 | 0.8477 | 0.8097 |

**Table 2** Average RMSD and bias (italic values) for estimating $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\rho$ when $N = 500$, $K = 40$

| | True $\rho = 0.2$ | | | | True $\rho = 0.5$ | | | | True $\rho = 0.8$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{\theta}_2$ | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 |
| $\boldsymbol{\alpha}_1$ | 0.1109 | 0.1214 | 0.1187 | 0.1204 | 0.2191 | 0.5224 | 0.4451 | 0.5294 | 0.1084 | 0.1446 | 0.1452 | 0.1478 |
| | *0.0670* | *0.0462* | *0.0467* | *0.0478* | *0.1142* | *0.1249* | *0.1165* | *0.1312* | *0.0590* | *0.0801* | *0.0806* | *0.0828* |
| $\boldsymbol{\alpha}_2$ | 0.1076 | 0.1088 | 0.1103 | 0.1094 | 0.1516 | 0.1365 | 0.1372 | 0.1336 | 0.0992 | 0.1261 | 0.1272 | 0.1289 |
| | *0.0357* | *0.0333* | *0.0369* | *0.0352* | *0.0883* | *0.0575* | *0.0655* | *0.0580* | *0.0299* | *0.0678* | *0.0692* | *0.0696* |
| $\boldsymbol{\delta}_1$ | 0.1266 | 0.1333 | 0.1347 | 0.1382 | 0.1703 | 0.2246 | 0.2333 | 0.2228 | 0.1213 | 0.1235 | 0.1281 | 0.1228 |
| | *−0.0417* | *−0.0549* | *−0.0559* | *−0.0569* | *−0.1217* | *−0.1610* | *−0.1696* | *−0.1608* | *−0.0414* | *−0.0419* | *−0.0532* | *−0.0502* |
| $\boldsymbol{\delta}_2$ | 0.1183 | 0.1259 | 0.1314 | 0.1264 | 0.1577 | 0.2125 | 0.2210 | 0.2110 | 0.1104 | 0.1152 | 0.1188 | 0.1161 |
| | *−0.0302* | *−0.0365* | *−0.0369* | *−0.0370* | *−0.1066* | *−0.1463* | *−0.1542* | *−0.1456* | *−0.0277* | *−0.0233* | *−0.0331* | *−0.0298* |
| $\rho_{12}$ | 0.0036 | 0.0169 | 0.0157 | 0.0174 | 0.0014 | 0.0076 | 0.0075 | 0.0073 | 0.0007 | 0.0116 | 0.0118 | 0.0118 |
| | *0.0092* | *0.0506* | *0.0505* | *0.0524* | *0.0019* | *0.0845* | *0.0834* | *0.0825* | *−0.0072* | *0.1074* | *0.1082* | *0.1079* |
| $r(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1)$ | 0.9224 | 0.9150 | 0.9157 | 0.9145 | 0.9325 | 0.9430 | 0.9431 | 0.9436 | 0.9467 | 0.9458 | 0.9448 | 0.9451 |
| $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ | 0.9422 | 0.8221 | 0.8518 | 0.8114 | 0.9448 | 0.8492 | 0.8690 | 0.8372 | 0.9409 | 0.8307 | 0.8508 | 0.8285 |

**Table 3** Average RMSD and bias (italic values) for estimating $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\rho$ when $N = 1000$, $K = 20$

| | True $\rho = 0.2$ | | | | True $\rho = 0.5$ | | | | True $\rho = 0.8$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{\theta}_2$ | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 |
| $\boldsymbol{\alpha}_1$ | 0.0680 | 0.0724 | 0.0739 | 0.0724 | 0.0720 | 0.0751 | 0.0770 | 0.0755 | 0.0709 | 0.0850 | 0.0854 | 0.0842 |
| | *0.0188* | *0.0213* | *0.0220* | *0.0207* | *0.0193* | *0.0200* | *0.0195* | *0.0190* | *0.0193* | *0.0426* | *0.0423* | *0.0419* |
| $\boldsymbol{\alpha}_2$ | 0.0753 | 0.1506 | 0.1499 | 0.1492 | 0.0832 | 0.1469 | 0.1451 | 0.1442 | 0.0705 | 0.1511 | 0.1503 | 0.1499 |
| | *0.0176* | *0.0052* | *0.0060* | *0.0077* | *0.0277* | *0.0121* | *0.0158* | *0.0168* | *0.0219* | *−0.0006* | *−0.0004* | *0.0001* |
| $\boldsymbol{\delta}_1$ | 0.0582 | 0.0612 | 0.0628 | 0.0632 | 0.0625 | 0.0668 | 0.0697 | 0.0694 | 0.0698 | 0.0703 | 0.0725 | 0.0717 |
| | *−0.0128* | *−0.0125* | *−0.0138* | *−0.0154* | *−0.0233* | *−0.0250* | *−0.0236* | *−0.0267* | *−0.0259* | *−0.0288* | *−0.0288* | *−0.0308* |
| $\boldsymbol{\delta}_2$ | 0.0659 | 0.0660 | 0.0679 | 0.0666 | 0.0618 | 0.0677 | 0.0689 | 0.0685 | 0.0705 | 0.0702 | 0.0714 | 0.0705 |
| | *−0.0011* | *−0.0089* | *−0.0099* | *−0.0100* | *−0.0112* | *−0.0179* | *−0.0161* | *−0.0190* | *−0.0131* | *−0.0232* | *−0.0228* | *−0.0249* |
| $\rho_{12}$ | 0.0011 | 0.0013 | 0.0013 | 0.0013 | 0.0010 | 0.0043 | 0.0044 | 0.0044 | 0.0004 | 0.0112 | 0.0112 | 0.0113 |
| | *−0.0070* | *0.0251* | *0.0252* | *0.0253* | *−0.0123* | *0.0601* | *0.0606* | *0.0608* | *−0.0130* | *0.1056* | *0.1054* | *0.1058* |
| $r(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1)$ | 0.9148 | 0.9139 | 0.9137 | 0.9139 | 0.9294 | 0.9200 | 0.9198 | 0.9200 | 0.9487 | 0.9426 | 0.9424 | 0.9426 |
| $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ | 0.9111 | 0.7875 | 0.8362 | 0.7811 | 0.9235 | 0.8024 | 0.8469 | 0.7958 | 0.9320 | 0.8199 | 0.8489 | 0.8181 |

**Table 4** Average RMSD and bias (italic values) for estimating $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\rho$ when $N = 1000$, $K = 40$

| | True $\rho = 0.2$ | | | | True $\rho = 0.5$ | | | | True $\rho = 0.8$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{\theta}_2$ | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 | Dist. 1 | Dist. 2 | Dist. 3 | Dist. 4 |
| $\boldsymbol{\alpha}_1$ | 0.0582 | 0.0673 | 0.0668 | 0.0666 | 0.0674 | 0.0686 | 0.0679 | 0.0679 | 0.0678 | 0.0905 | 0.0897 | 0.0906 |
| | *0.0239* | *0.0286* | *0.0288* | *0.0281* | *0.0243* | *0.0318* | *0.0317* | *0.0314* | *0.0249* | *0.0585* | *0.0582* | *0.0582* |
| $\boldsymbol{\alpha}_2$ | 0.0737 | 0.0740 | 0.0737 | 0.0742 | 0.0737 | 0.0765 | 0.0761 | 0.0764 | 0.0725 | 0.1060 | 0.1060 | 0.1055 |
| | *0.0218* | *0.0230* | *0.0225* | *0.0229* | *0.0288* | *0.0273* | *0.0269* | *0.0274* | *0.0292* | *0.0626* | *0.0619* | *0.0620* |
| $\boldsymbol{\delta}_1$ | 0.0718 | 0.0800 | 0.0795 | 0.0808 | 0.0929 | 0.0931 | 0.0932 | 0.0942 | 0.0937 | 0.1102 | 0.1024 | 0.1067 |
| | *−0.0435* | *−0.0437* | *−0.0426* | *−0.0449* | *−0.0549* | *−0.0592* | *−0.0542* | *−0.0605* | *−0.0599* | *−0.0752* | *−0.0662* | *−0.0722* |
| $\boldsymbol{\delta}_2$ | 0.0689 | 0.0761 | 0.0756 | 0.0762 | 0.0815 | 0.0864 | 0.0891 | 0.0889 | 0.0889 | 0.0933 | 0.0968 | 0.1010 |
| | *−0.0357* | *−0.0361* | *−0.0352* | *−0.0376* | *−0.0485* | *−0.0499* | *−0.0453* | *−0.0521* | *−0.0514* | *−0.0652* | *−0.0571* | *−0.0635* |
| $\rho_{12}$ | 0.0015 | 0.0015 | 0.0015 | 0.0015 | 0.0009 | 0.0055 | 0.0055 | 0.0056 | 0.0002 | 0.0125 | 0.0125 | 0.0125 |
| | *−0.0008* | *0.0273* | *0.0275* | *0.0275* | *−0.0004* | *0.0708* | *0.0708* | *0.0711* | *−0.0021* | *0.1115* | *0.1115* | *0.1116* |
| $r(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1)$ | 0.9537 | 0.9532 | 0.9533 | 0.9530 | 0.9575 | 0.9545 | 0.9544 | 0.9542 | 0.9575 | 0.9624 | 0.9624 | 0.9621 |
| $r(\boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_2)$ | 0.9527 | 0.8539 | 0.8769 | 0.8338 | 0.9516 | 0.8567 | 0.8771 | 0.8388 | 0.9570 | 0.8609 | 0.8721 | 0.8508 |

## 4 Conclusion and Discussion

In general, with the use of Monte Carlo simulations, this study demonstrates that departure from normal distributions for the latent traits in the multi-unidimensional GRM does affect the accuracy of its parameter recovery. This is in line with findings from previous studies with unidimensional IRT models (e.g., Sass et al. 2008; Reise and Revicki 2014). Specifically, what we found in our study are that skewed distributions would affect more on the accuracy in estimating the item step parameters and that kurtotic distributions affect the estimation of person parameters. In situations where not all latent traits are normally distributed (such as what was considered in the simulation study), the non-normal shape associated with a few latent traits would affect the estimation of parameters in other dimensions when the intertrait correlation is moderate to high. As non-normal latent trait distributions are common in many polytomous response items, and examples of such instruments include mental tests, business satisfaction, cross-cultural differences, etc., one needs to be aware of the shapes of latent trait distributions before fitting the model to actual data. However, such information may not always be available in practice. It is hence important to find alternate solutions, such as using a more robust estimation method or a non-normal prior distribution. In addition, this study shows that increased sample size and/or test length can help improve the estimation of the multi-unidimensional GRM parameters. This finding not only confirms results from previous studies dealing with normal latent trait(s) (e.g., Linacre 2002; Sheng 2010; Kuo and Sheng 2015; Wollack et al. 2002) but also extends to situations where the latent traits are not normal. One may consider reducing the effect of non-normality by increasing sample size/test length under the non-normal conditions. The minimum number of persons/items necessary to reach a desired level of accuracy can be an interesting study that requires further investigation.

This study focuses on Likert scale items involving three scales, and therefore two threshold parameters need to be estimated for each item. Further study can evaluate the estimation of these procedures using items with more than three scales or with different numbers of scales. In addition, this study investigates the effects of non-normal latent traits using the HwG estimation method. Further study can include other estimation techniques, such as marginal maximum likelihood (Bock and Aitkin 1981) and Metropolis-Hastings Robbins-Monro (Cai 2010a,b). Lastly, the simulation study adopted 25 replications due to the computational expense of the MCMC procedures. Further studies can consider more replications to achieve a better accuracy.

## References

J.H. Albert, S. Chib, Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**, 669–679 (1993)

R.D. Bock, M. Aitkin, Marginal maximum likelihood estimation of item parameters: applications of an EM algorithm. Psychometrika **46**, 443–459 (1981)

R.D. Buchanan, The development of the Minnesota multiphasic personality inventory. J. Hist. Behav. Sci. **30**, 148–161 (1994)
L. Cai, High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. Psychometrika **75**, 33–57 (2010a)
L. Cai, Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. J. Educ. Behav. Stat. **35**, 307–335 (2010b)
M.K. Cowels, Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Stat. Comput. **6**, 101–111 (1996)
C.G. Ferero, A. Maydeu-Olivares, Estimation of IRT graded response models: limited versus full information methods. Psychol. Methods **14**, 275–299 (2009)
Z.H. Fu, J. Tao, N.Z. Shi, Bayesian estimation of the multidimensional graded response model with nonignorable missing data. J. Stat. Comput. Simul. **80**, 1237–1252 (2010)
S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
M. Harwell, C.A. Stone, T.-C. Hsu, L. Kirisci, Monte carlo studies in item response theory. Appl. Psychol. Meas. **20**, 101–125 (1996)
W. Hastings, Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
L. Kirisci, T. Hsu, L. Yu, Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. Appl. Psychol. Meas. **25**, 146–162 (2001)
J. Koran, T.C. Headrick, T.-C. Kuo, Simulating univariate and multivariate nonnormal distributions through the method of percentiles. Multivar. Behav. Res. **50**, 1–17 (2015)
T.C. Kuo, Y. Sheng, Bayesian estimation of a multi-unidimensional graded response IRT model. Behaviormetrika **42**, 79–94 (2015)
R. Likert, A technique for the measurement of attitudes. Arch. Psychol. **22**, 5–55 (1932)
J.M. Linacre, Optimizing rating scale category effectiveness. J. Appl. Meas. **3**, 85–106 (2002)
F.M. Lord, *Applications of Item Response Theory to Practical Testing Problems* (Lawrence Erlbaum, Hillsdale, 1980)
R.P. McDonald, *Test Theory: A Unified Approach* (Lawrence Erlbaum, Mahwah, 1999)
N. Metropolis, S. Ulam, The Monte Carlo method. J. Am. Stat. Assoc. **44**, 335–341 (1949)
M. Reckase, The past and future of multidimensional item response theory. Appl. Psychol. Meas. **21**, 25–36 (1997)
M. Reckase, *Multidimensional Item Response Theory* (Springer, New York, 2009)
S. Reise, D. Revicki, *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. Multivariate Applications Book Series (Routledge, New York, 2014)
V.J. Rubio, D. Aguado, P.M. Hontangas, J.M. Hernandez, Psychometric properties of an emotional adjustment measure. Eur. J. Psychol. Assess. **23**, 39–46 (2007)
F. Samejima, Estimation of latent ability using a response pattern of graded scores. Psychometrika **35**, 139–139 (1969)
J.R.S. Santo, C.L.N. Azevedo, H. Bolfarine, A multiple group item response theory model with centred skew normal latent trait distributions under a bayesian framework. J. Appl. Stat. **40**, 2129–2149 (2013)
D.A. Sass, T.A. Schmitt, C.M. Walker, Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. Appl. Meas. Educ. **21**, 65–88 (2008)
Y. Sheng, A sensitivity analysis of gibbs sampling for 3PNO IRT models: effects of prior specifspecific on parameter estimates. Behaviormetrika **37**, 87–110 (2010)
Y. Sheng, T.C. Headrick, A Gibbs sampler for the multidimensional item response model. ISRN Appl. Math. **2012**, 14pp. (2012)
Y. Sheng, C.K. Wikle, Comparing multiunidimensional and unidimensional item response theroy models. Educ. Psychol. Meas. **67**, 899–919 (2007)
Y. Sheng, A MATLAB package for Markov chain Monte Carlo with a multi-unidimensional IRT model. J. Stat. Softw. **28**, 1–20 (2008)
J.A. Wollack, D.M. Bolt, A.S. Cohen, Y.S. Lee, Recovery of item parameters in the nominal response model: a comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. Appl. Psychol. Meas. **26**, 339–352 (2002)