**Examining the Effects of Non-normality on Parameter Estimation in Bifactor Graded Response Models**

**Abstract**

Psychometric models such as item response theory (IRT) frequently presume a normal distribution of latent abilities and traits. However, violations of normality assumption often exist in psychology and psychiatric research areas. Despite extensive research on the effects of non-normality in IRT and other psychometric models, little attention has been given to its impact on bifactor IRT models, particularly with polytomous data. This study focuses on the issue of non-normality in parameter estimation in the bifactor-graded response model (Bifactor-GRM). The results show that non-normality in the general factor significantly affects the accuracy of estimating general-factor discrimination and thresholds/locations but not specific-factor discrimination. Meanwhile, non-normality in the specific factor has a negligible effect on both general- and specific-factor item parameter estimation. However, we can minimize the effect of non-normality by increasing sample sizes and item numbers. Regarding person parameter estimation, the maximum a posteriori (MAP) estimation generally outperforms the marginal maximum likelihood (ML), especially when facing a violation of normality in population abilities or traits. To lessen the bias of the ML method, researchers also can increase sample sizes and item numbers to enhance accuracy.

*Keywords:* Bifactor model, graded response model, non-normality

**Introduction**

In psychology and psychiatric research areas, it is common to encounter a latent construct that is positively skewed. For example, most people are at the normal end of a psychiatric disorder spectrum, while a smaller number of individuals spread out along the continuum of the disorder end. However, many latent variable approaches, such as item response theory (IRT) and factor analytic methods, assume the normality of the latent trait of interest. The impact of the non-normality on parameter estimation of latent variable approaches has been attracting researchers’ attention (e.g., Wang et al., 2018). Previous research has primarily focused on exploring the effects of non-normality on structural equation modeling (SEM) (Finch et al., 1997; Lai, 2018; Lei & Lomax, 2005; Maydeu-Olivares, 2017; Olsson et al., 2000; Ory & Mokhtarian, 2010), and confirmatory factor analysis (CFA) (Curran et al., 1996; Hutchinson & Olmos, 1998; Savalei, 2008). There has been relatively less research in investigating non-normality in the context of item response theory (IRT) models (Svetina et al., 2017; Woods, 2014), particularly using the bifactor IRT model. The Bifactor model has been gaining popularity in psychological and other social sciences because of its flexibility in incorporating a general factor and some specific factors for the multidimensional latent factors. To the best of our knowledge, no previous study has examined the impact of non-normality on bifactor models’ parameter estimation. This study will focus on the impact of the violation of the assumption of normality in the bifactor model with the graded response data. It is an extension of previous studies focused on unidimensional IRT models (DeMars, 2012; Sen et al., 2016) and multidimensional IRT models (Svetina et al., 2017; Wang et al., 2018; Woods, 2014).

Compared to previous research studies designed for normality violation in unidimensional or multidimensional models, the current study uses bifactor graded response model (Bifactor-GRM) to examine how the skewness and kurtosis of the general factor and specific factors affect the recovery of parameters, including item parameters and person ability estimates. The design factors included the severity of skewness of the general factor and specification factors, sample size, the number of factors, and the number of items per factor.

Most commercial software and open-source packages offer one or more than one estimation method to estimate the parameters of models, but most of them are based on the normal distribution. For person parameter estimation, the marginal maximum likelihood (ML) method is the most widely used approach for estimating item parameters, while maximum a posteriori (MAP) estimation has been shown to achieve more accurate estimation with fewer items but also requires the assumption of normality of person parameters (Brown, 2015). In this study, MAP and ML estimation are used to estimate the parameters of the bifactor IRT model.

**Bifactor Grade Response Model**

The Bifactor-GRM is an extension of the conventional GRM and is a part of IRT models. In a Bifactor-GRM, items are allowed to load onto a general factor (akin to a general ability or trait in the respondent) and one or more specific or group factors (specific abilities or traits) (Reise et al., 2010). The probability that an examinee’s response falls at or above a particular ordered category given *θ*.

In which P is the probability to provide a response equal to k or greater given a person's location on general factor (*G*) and a specific trait (*S*), category k's item-intercept as defined , and the conditional item discrimination parameter on *G* () and on *S* (). The person parameter represents person *i*'s location (ability or trait) on *G*, whereas represents person *i*’s location on *S*. For each person there are several specific trait scores equivalent to the number of specific traits defining the model (Toland et al., 2017).

Based on Equation (1), the category response functions, which indicate the probability of responding to a particular category given *θ*, can be calculated by subtraction of adjacent boundary functions,

**Method**

**Design Factors**

This study conducts a Monte Carlo simulation to assess the effects of non-normality on latent trait distributions within a bifactor model, which is a common framework in psychometric and educational assessments. The model accounts for both a general factor and various specific factors that describe individual abilities. Furthermore, the research incorporates four-category polytomous items, which are frequently used in these fields. Drawing on methodologies established in prior studies (Auné, 2020; Mao, 2022; Rijmen, 2011; Svetina et al., 2017; Wang et al., 2018), it explores how deviations from normality in these latent traits affect model outcomes. In a bifactor model, individuals are characterized by a single general factor (*θg*) and several specific factors (*θsk*), where k represents the number of specific factors. The simulation aims to elucidate the impact of latent trait non-normality on model performance, providing insights that are critical for both theoretical understanding and practical application in psychological and educational evaluation.

**Table 1.**

*Simulation Design*

|  |  |  |
| --- | --- | --- |
| **Design factors** | **Levels** | **Values of Levels** |
| Sample size (*N*) | 3 | *N* = 250, 500, 1000 |
| Number of Item per Factor (*Is*) | 3 | *Is* = 5, 10, 20 |
| Number of Specific Factor (*Fs*) | 2 | *Fs* = 2, 4 |
| Non-normality on general factor (*Normg*) | 2 | Normal : Skew = 0, Kurt = 0,  Non-normal : Skew = 2, Kurt = 7 |
| Non-normality on general factor (*Norms*) | 2 | Normal : Skew = 0, Kurt = 0,  Non-normal : Skew = 2, Kurt = 7 |

Table 1 presents a summary of the manipulated factors in our simulation study, including sample sizes (*N* = 250, 500, 1000), the number of items per factor (*Is* = 5, 10, 20), and either two or four specific factors (*Fs* = 2, 4). Our study simulates conditions of non-normality for both the general factor (*Normg* = Normal, Non-normal) and the specific factor (*Norms* = Normal, Non-normal). The normal condition is characterized by a skewness of 0 and a kurtosis of 0, while the non-normal condition is characterized by a skewness of 2 and a kurtosis of 7, both of which were used in previous research (Curran et al., 1996; Wang et al., 2018).

The experimental design comprehensively intersects these manipulated variables, generating 72 unique simulation conditions. Each condition underwent 1000 replications through the "SimMultiCorrData" package in R (Fialkowski, 2018), ensuring robust and detailed analysis of the bifactor model's performance under varying degrees of non-normality.

**Item parameter**

In psychological and psychiatric research, the general factor discrimination is usually positive and falls within the range of 1.1 to 2.8 (Atkinson, 2018; Auné, 2020; Berkeljon, 2012; Raines, 2015). Previous studies have consistently shown that specific factor discriminations are typically smaller than the general factor, ranging from 0 to 1.5 (Wang et al., 2018). In the bifactor model, the general factor and specific factor are considered independent, with no correlation between them. In this study, the discrimination values for the general factor are set to range from 1.1 to 2.8, while the discrimination values for the specific factors are established within the range of 0 to 1.5.

Item difficulty values can theoretically range from negative infinity to positive infinity, but in practice, they typically vary from -2 to +2 (Hambleton, 1993; Hambleton & Swaminathan, 1985). Psychological and psychiatric tests often use a four-point Likert scale to measure latent traits or personalities (Auné et al., 2020; Rijmen,2011). According to Wang (2018), this study generated normally distributed thresholds, b1[−2, −0.67], b2[−0.67, 0.67], and b3[0.67, 2], for three thresholds (locations) to distinguish the possibilities of choosing each item.

**Person parameter**

The values for skewness and kurtosis between -2 and +2 are considered acceptable for assuming normality (George & Mallery, 2010). Hair et al. (2010) and Bryne (2010) argued that data is normal if skewness is between ‐2 to +2 and kurtosis is between ‐7 to +7. Thus, we simulate two levels of non-normality, normality (skewness: 0, kurtosis: 0), and non-normality (skewness: 2, kurtosis: 7). There were four combinations of normality status for the general factor and specific factors. In this study, we employed the Fleishman method to generate nonnormal distributions; this technique involves manipulating a normally distributed random variable using a cubic polynomial, thereby adjusting skewness and kurtosis through modification of the polynomial's coefficients (Fleishman, 1978). All latent traits on specific factors (θs) are set equally.

**Estimation**

The item parameters in this study were estimated using the "bfactor()" function from the R package "mirt", limited in 6000 iterations. For estimating the person ability parameters, two estimation methods, namely, maximum a posteriori (MAP) and maximum likelihood (ML), were utilized. Within the R package "mirt," the estimation of person ability parameters involved utilizing the "fscores()" function. In this package, the thresholds or locations are calculated as cjk, as described in Equation (1).

***Evaluation criteria***

The accuracy of parameter recovery in this study is assessed through the calculation of bias, root mean squared error (RMSE), and Pearson correlations (only for person ability). These measures are calculated for both the two discrimination parameters, the three boundary parameters, and two personal parameters.

**Bias.** The relative bias is estimated for all the parameters of model, including item parameters (ag, as, c1, c2, c3) and personal parameter (θg, θs) as,

where is the estimated parameters () across valid replications and is the true parameters (agj, asj, c1j, c2j, c3j , θgj, θsj,). In the , j represents the item number, ranging from 1 to J. The total number of items J is computed by multiplying the number of items in each specific factor by the number of specific factors. For each condition, a total of 500 replications are carried out, denoted as R in equation (3).

**RMSE**. The RMSE is estimated for all the parameters of model, including item parameter (ag, as, c1, c2, c3) and personal parameter (θg, θs) as,

where refers to the same as mentioned equation (3).

To determine the effect of the design factors on the outcome variables, we conducted a factorial analysis of variance (ANOVA) with effect size (η2) computed to gauge the contribution of all the design factors and their interaction. Note that only the practically significant design factors and their interactions are considered as salient effect based on J. Cohen’s (2013) small effect size of 0.01, moderate effect size of 0.06, and large effect size of 0.14.

**Results**

**Item Parameter Estimation**

In this section, we only looked at ANOVA effect size (*η²*) exceeding 0.06, to get a good idea of how non-normality and other design factors affected estimating item parameters for bias and RMSE.

**Table 2**

*Generalized Eta Squared (η²)* *for Item Parameter Estimates*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Source of Variation |  | Bias |  |  | RMSE |  |
| *ag* | *as* | *c* | *ag* | *as* | *c* |
| *Normg* | .498 |  | .167 |  |  |  |
| *Is* |  |  |  | .098 | .199 | .115 |
| *N* | .064 |  |  | .109 | .136 | .282 |
| *Is× N* |  |  |  |  | .082 |  |

***Note.*** *Normg*: Non-normality in the general factor; *Is*: Item number per specific factor; *N*: Sample size; *Is× N*: Interaction between *Is* and *N*;*ag*: discrimination of the general factor; *as*: average discrimination of specific factors; *c*: average value of thresholds or locations (*c1*, *c2*, and *c3*).

According to Table 2, non-normality in the general factor (*Normg*) primarily affected the bias in estimating discrimination of general factor (*ag*) and thresholds or locations (*c*), with *η²* = 0.498 and 0.167, respectively. Regarding RMSE, *Normg* had a negligible effect on all item parameter estimations, while item number per specific factor (*Is*) and sample size (*N*) had a significant influence. Additionally, the only interaction between *Is* and *N* (0.082) had an effect size over the threshold, significantly influencing RMSE of estimating the discrimination on specific factors (*as*). We will explain this interaction effect in the following section.

**Table 3**

*Means and SDs of Bias and RMSE in Item Parameter Estimation Across Different Conditions*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable and Condition | | *ag* | *as* | *c1* | *c2* | *c3* |
| **Bias** |  |  |  |  |  |  |
| *Normg* | Normal | **0.03(0.28)** | -0.01(0.30) | **-0.03(0.23)** | **-0.01(0.19)** | **0.01(0.23)** |
|  | Non-normal | **-0.19(0.29)** | 0.00(0.32) | **-0.10(0.24)** | **-0.10(0.18)** | **-0.07(0.21)** |
| *Is* | 5 | -0.04(0.56) | 0.04(0.69) | -0.12(0.39) | -0.08(0.22) | -0.01(0.36) |
|  | 10 | -0.09(0.25) | -0.01(0.22) | -0.07(0.20) | -0.06(0.18) | -0.04(0.20) |
|  | 20 | -0.09(0.22) | -0.01(0.16) | -0.05(0.20) | -0.05(0.18) | -0.03(0.20) |
| *N* | 250 | **-0.05(0.42)** | 0.01(0.46) | -0.09(0.33) | -0.06(0.25) | -0.01(0.31) |
|  | 500 | **-0.09(0.24)** | -0.01(0.24) | -0.06(0.20) | -0.05(0.17) | -0.03(0.20) |
|  | 1000 | **-0.10(0.18)** | -0.02(0.14) | -0.06(0.14) | -0.06(0.13) | -0.04(0.14) |
| ***RMSE*** |  |  |  |  |  |  |
| *Normg* | Normal | 0.28(2.01) | 0.30(2.22) | 0.23(1.20) | 0.19(0.61) | 0.24(1.31) |
|  | Non-normal | 0.34(2.97) | 0.32(3.98) | 0.26(2.07) | 0.20(0.37) | 0.22(0.80) |
| *Is* | 5 | **0.57(4.26)** | **0.69(5.57)** | **0.41(2.90)** | **0.23(0.86)** | **0.36(1.85)** |
|  | 10 | **0.26(0.57)** | **0.22(0.52)** | **0.22(0.40)** | **0.19(0.26)** | **0.20(0.36)** |
|  | 20 | **0.23(0.30)** | **0.16(0.23)** | **0.21(0.28)** | **0.19(0.25)** | **0.20(0.26)** |
| *N* | 250 | **0.43(3.45)** | **0.46(4.51)** | **0.35(2.35)** | **0.25(0.70)** | **0.31(1.50)** |
|  | 500 | **0.26(0.50)** | **0.24(0.61)** | **0.21(0.42)** | **0.18(0.23)** | **0.20(0.37)** |
|  | 1000 | **0.20(0.28)** | **0.14(0.27)** | **0.15(0.21)** | **0.14(0.17)** | **0.14(0.19)** |

***Note.*** *Normg*: Non-normality on general factor; *Is*: Number of Item on specific factors; *N*: Sample size;*ag*: discrimination of general factor; *as*: average discrimination of specific factors; *c*: average value of thresholds or locations (*c1*, *c2*, and *c3*).

To get a better understanding of the main effects of simulation factors, we conducted Table 3 to address all means and standard deviations (SDs) of bias and RMSE in item parameter estimation across all simulation conditions. The table shows the bias and RMSE values for five item parameters (ag, as, c1, c2, c3) when the conditions are different. It compares the effects of data normality (Normg), the number of items (Is), and the sample size (N). Notably, when the general-factor ability is non-normally distributed (Normg = Non-normal), it generally exhibits a negative and larger deviation from zero bias of estimation on four item parameters (*ag*, *c1*, *c2*, and *c3*), indicating decreasing accurate estimation. Meanwhile, non-normality on the general factor does not impact the estimation on as*.* In terms of RMSE, non-normality on the general factor (Normg = Non-normal) doesn't have a big impact on estimation for any of the five item parameters (ag, as, c1, c2, c3).

Additionally, **item numbers per specific factor (Is)**, increasing from 5 to 20, consistently reduce both bias and RMSE, suggesting that more items improve estimation stability. Similarly, **a sample size (N)** increasing from 250 to 1000 leads to lower bias and RMSE, too. At N = 1000, the most stable and accurate results are observed. Overall, the results suggest that non-normally distributed ability on the general factor, larger item sets, and bigger sample sizes all contribute to better estimation accuracy.

**Figure 1**

*Interaction Effect on RMSE of Specific-factor Discrimination (as)*

A graph of a sample size and a sample size

Description automatically generated with medium confidence

***Note.*** *as*: specific-factor discrimination parameter in bifactor GRM model; RMSE: Root Mean Square Error; General Factor: the normality of the simulated general-factor latent ability distribution.

Figure 1 shows that the RMSE for estimating specific-factor discrimination parameter (*αs*) decreases as both sample size and the number of items per specific factor increase. The existence of non-normality on general-factor latent ability (*θg*) had little impact on RMSE, suggesting that general factor non-normality does not substantially affect specific-factor discrimination (*αs*) estimation. Thus, increasing sample size and the number of items per specific factor significantly improves the accuracy of specific-factor discrimination (*αs*) estimation, as measured by RMSE.

**Person Parameter Estimation**

**Table 4**

*Generalized Eta Squared (**η2) for Person Parameter Estimates*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variation | Bias | | RMSE | | |
| *θg* | *θs* |  | *θg* | *θs* |
| *Normg* | **.163** | .003 |  | .039 | **.082** |
| *SF* | .001 | .047 |  | .009 | .038 |
| *Fs* | .004 | .036 |  | **.539** | **.657** |
| *Is* | .012 | **.092** |  | **.742** | **.880** |
| *N* | .002 | .000 |  | .005 | .016 |
| *Method* | .042 | .020 |  | **.856** | **.967** |
| *SF× Is* | .000 | .004 |  | **.117** | **.168** |
| *Is× N* | .000 | .002 |  | .041 | .014 |
| *Is× Normg* | .004 | **.091** |  | .002 | .005 |
| *Fs× Method* | .000 | .028 |  | **.264** | **.179** |
| *Is× Method* | .001 | **.093** |  | **.589** | **.826** |
| *Normg × Method* | .031 | .000 |  | .006 | **.074** |
| *Fs× N× Normg* | .000 | .001 |  | .001 | .009 |
| *Is× N× SF* | .004 | .002 |  | **.072** | .011 |
| *Fs× Is× Method* | .001 | .004 |  | **.079** | **.085** |
| *Is× Normg × Method* | .003 | **.093** |  | .001 | .005 |
| *Fs× Is× N× Normg* | .003 | .003 |  | .011 | .029 |
| *Fs: Is: N: Normg: Norms* | .001 | .001 |  | .049 | .003 |

***Note.*** *Normg*: Non-normality of the general factor; *SF*: Non-normality of the specific factor; *Fs*: Number of specific factors; *Is*: Number of items per specific factor; *N*: Sample size; *Method*: Person parameter estimation method (MAP or ML); *θg*: Estimated general-factor ability; *θs*: Estimated specific-factor ability. Interactive effects are noted by multiplication sign (*×*).

Table 4 reveals several critical insights in examining the influence of non-normality, integrated by number of specific factors (*Fs*), the number of items per specific factor (*Is*), sample size (*N*), and the algorithm for person parameter estimation. In terms of bias, general-factor non-normality (*Normg*) accounted for a substantial proportion of variance in general-factor ability (*θg*) estimation (*η2* = .163), while the number of items per specific factor (*Is*) accounted for a smaller proportion of variance in specific-factor ability (*θs*) estimation (*η2* = .092). Regarding RMSE, we can find more meaningful impacts. Specifically, the estimation algorithm had the largest effect on both general-factor ability (*θg*) and specific-factor ability (*θs*), with *η2* values of .856 and .967, respectively. This highlights the critical importance of choosing between MAP and ML estimation for accurate person parameter recovery.

Table 4 also reveals interaction effects, but we only discuss two-way interactions here and leave three-way interactions discussed and plotted in the next section. There were only two two-way interactions found to influence specific-factor ability (*θs*) with respect to bias. These were the interaction between the number of items per specific factor (*Is*) and general-factor non-normality (*Normg*), and the interaction between *Is* and the estimation method (*Method*). Regarding RMSE, several two-way interactions accounted for variance in both *θg* and *θs*. For *θg*, three two-way interactions were found, the interaction between specific-factor non-normality and the number of items per specific factor (*SF× Is*, *η2* = .117), the number of specific factors and estimation method (*Fs× Method*, *η2* = .264), and non-normality of the specific factor and estimation method (*Is× Method*, *η2* = .589). Meanwhile, for *θs*, four two-way interactions were addressed, the interaction between specific-factor non-normality and the number of items per specific factor (*SF× Is*, *η2* = .168), the number of specific factors and estimation method (*Fs× Method*, *η2* = .179), non-normality of the specific factor and estimation method (*Is× Method*, *η2* = .826), and non-normality of the general factor and estimation method (*Normg× Method*, *η2* = .074). Although the main effect of specific-factor non-normality (*SF*) was not statistically significant, SF was a key component in several significant two-way interaction effects. The main effect of general-factor non-normality (*Normg*), however, was comparatively small.

In conclusion, general-factor non-normality (*Normg*) had a substantial impact on bias in general-factor ability (*θg*) estimation, while only the number of items per specific factor (*Is*) had a smaller but notable role in specific-factor ability (*θs*) estimation. Notably, the choice of estimation algorithm (MAP or ML) was the most influential factor, significantly affecting both *θg* and *θs* in terms of RMSE. In several identified two-way interactions, the number of items per specific factor (*Is*) and the estimation method pronounced major effects on RMSE, and bias of *θs* estimation. Unexpectedly, sample size (*N*) exhibited minor impact in both one-way and two-way interactions. Unexpectedly, sample size (*N*) exhibited minor impact in both one-way and two-way interactions. These findinNorms underscore the importance of carefully selecting estimation methods and considering factor number and item number in non-normal bi-factor model’s person parameter estimation.

**Table 5**

*Bias and RMSE in Person Parameter Estimation Across Different Conditions*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable and Condition | | Bias | | RMSE | | |
| *θg* | *θs* | | *θg* | *θs* |
| *Normg* | Normal | 0.00(0.06) | 0.01(0.06) | | 0.50(0.32) | 1.07(0.67) |
|  | Non-N | 0.05(0.06) | 0.00(0.08) | | 0.53(0.33) | 1.01(0.63) |
| *Norms* | Normal | 0.03(0.06) | -.01(0.07) | | 0.52(0.34) | 1.02(0.65) |
|  | Non-N | 0.02(0.06) | 0.02(0.07) | | 0.51(0.32) | 1.06(0.65) |
| *Fs* | 2 | 0.03(0.06) | 0.00(0.08) | | 0.61(0.38) | 1.17(0.70) |
|  | 4 | 0.02(0.06) | 0.02(0.07) | | 0.42(0.23) | 0.90(0.56) |
| *Is* | 5 | 0.03(0.06) | -.02(0.09) | | 0.71(0.42) | 1.37(0.84) |
|  | 10 | 0.02(0.06) | 0.01(0.06) | | 0.48(0.23) | 1.01(0.52) |
|  | 20 | 0.02(0.06) | 0.02(0.05) | | 0.35(0.14) | 0.74(0.31) |
| *N* | 250 | 0.02(0.07) | 0.01(0.08) | | 0.51(0.30) | 1.02(0.63) |
|  | 500 | 0.02(0.06) | 0.01(0.07) | | 0.52(0.34) | 1.05(0.66) |
|  | 1000 | 0.03(0.05) | 0.00(0.06) | | 0.52(0.33) | 1.04(0.66) |
| *Method* | MAP | 0.01(0.05) | 0.00(0.03) | | 0.30(0.07) | 0.52(0.11) |
|  | ML | 0.04(0.06) | 0.01(0.10) | | 0.73(0.34) | 1.56(0.54) |

***Note.*** The table displays means and standard deviations (SDs, in parentheses) of bias and RMSE under various conditions in person parameter estimation. *Normg*: Non-normality of the general factor; *SF*: Non-normality of the specific factor; *Fs*: Number of specific factors; *Is*: Number of items per specific factor; *N*: Sample size; *Method*: Person parameter estimation method (MAP or ML); *θg*: Estimated general-factor ability; *θs*: Estimated specific-factor ability.

In Table 5, the person’s general-factor trait/ability(*θg*) shows small to positive bias across all conditions, with non-normal distribution conditions on general factor showing the most significant bias (0.05), compared to normal distribution condition (0.00), and ML method exhibits higher bias (0.04) than MAP method (0.01). Meanwhile, the person’s specific-factor traits/abilities(*θs*) are generally close to zero across all conditions, meaning estimation of *θs* is less affected by other factors.

Regarding RMSE, the *θg* consistently exhibited lower error rates compared to *θs* across various conditions. Notably, RMSE for *θg* decreased with an increased number of items and larger sample sizes, while sample size had minor impact. Meanwhile, the MAP method overperformance the ML method both in estimating the general-factor and the specific-factor abilities. Conversely, RMSE values for *θs* were consistently higher, suggesting greater variability and less precision in estimating person parameters. Like *θg*, the RMSE for *θs* also improved with more items and larger sample sizes, with the most significant enhancements observed when using the MAP method, in contrast to ML.

**Figure 2**

*Bias of Person General-factor Ability (𝜃g) Estimation*A graph of a number of items

Description automatically generated with medium confidence

***Note.*** MAP: The Maximum A Posteriori estimation; ML: the Maximum Likelihood estimation. Normality on general factor is depicted by the solid line, while non-normality on general factor is depicted by the dashed line.

Figure 2 compares the bias in general-factor ability (𝜃g) estimation between MAP and ML methods, highlighting the impact of non-normality distribution existing on general-factor ability. When using the MAP method, both normal (solid line) and non-normal (dashed line) distributions of general-factor ability result in minimal bias, which remains stable across varying numbers of items per specific item. In contrast, while the ML method also shows low bias in estimating general-factor ability, it exhibits a noticeable increase in bias as the number of items grows, from negative to positive, suggesting that ML is more sensitive to non-normal distribution of general-factor ability.

Overall, the results demonstrate that MAP provides more stable estimates under different distributional assumptions, whereas ML can become significantly biased in the presence of non-normality. Specifically, smaller numbers of items per specific factor led to notable negative bias, while larger numbers of items result in positive bias.

**Figure 3**

*RMSE of Person General-factor Ability (𝜃g) Estimation: interaction effect of Sample Size, Item Number, and Non-normality on Specific Factor*

A graph of different sizes and sizes of items

Description automatically generated with medium confidence

***Note.*** Normality on general factor is depicted by the solid line, while non-normality on general factor is depicted by the dashed line.

Figure 3 compares the RMSE in general-factor ability (*𝜃g*) estimation across sample size, item number per specific factor (*Is*), and non-normality on specific factor (*Norms*). Overall, only item number per specific factor had a noticeable effect on RMSE of person general-factor ability (*𝜃g*) estimation. This indicates that increasing the number of items within each specific factor significantly enhances the accuracy of estimating general-factor ability.

**Figure 4**

*RMSE of Person General-factor Ability (𝜃g) Estimation: interaction effect of Method, Item Number, and Specific Factor Number*

A graph of a number of items

Description automatically generated with medium confidence

***Note.*** MAP: The Maximum A Posteriori estimation; ML: the Maximum Likelihood estimation. The two specific factors condition is depicted by the solid line, while the four specific factors condition is depicted by the dashed line.

Figure 4 compares the RMSE in general-factor ability (*𝜃g*) estimation across methods, item numbers per specific factor (*Is*), and specific factor numbers (*Fs*). Overall, the MAP method outperforms the ML method, exhibiting lower and more stable RMSE values across varying item numbers. While a decrease in *Fs* and *Is* results in a slight reduction in RMSE under the MAP method, these factors have a more significant impact under the ML method. In conclusion, increasing both *Fs* and *Is* can notably diminish the performance difference between the MAP and ML methods.

**Figure 5**

*RMSE of Person General-factor Ability (𝜃s) Estimation: interaction effect of Method, Item Number, and Specific Factor Number*

A graph of a number of items

Description automatically generated with medium confidence

***Note.*** MAP: The Maximum A Posteriori estimation; ML: the Maximum Likelihood estimation. The two specific factors condition is depicted by the solid line, while the four specific factors condition is depicted by the dashed line.

Figure 5 addressed the RMSE in specific-factor ability (*𝜃s*) estimation across methods, item numbers per specific factor (*Is*), and specific factor numbers (*Fs*). Similar to the RMSE in general-factor ability (*𝜃g*) estimation, the MAP method outperforms the ML method. Both methods show that increasing *Fs* and *Is* lead to a reduction in RMSE under the MAP method. Additionally, the ML method exhibits extremely high RMSE, particularly under conditions with a low number of items per specific factor. This indicates that specific-factor ability (*𝜃s*) estimation is more sensitive to a small number of items within each specific factor.

**Discussions**

The non-normality in latent ability/trait distributions is a common psychological and educational assessment. As a frequently used model, the bifactor graded response model (Bifactor-GRM) might face a more complicated estimation problem when a non-normal distribution can happen on both the general factor and the specific factor. This section will address violations of the normality assumption in both item and person parameter estimation within the Bifactor-GRM and offer implications for both researchers and practitioners in psychometric areas.

Our results revealed that non-normality in the general factor predominantly influenced the bias in estimating the discrimination on the general factor (*ag*) and, to a lesser extent, the location parameters (*c1*, *c2*, and *c3*). In detail, general-factor non-normality resulted in increased negative bias in the discrimination parameter (*ag*) estimation. Regarding RMSE, general-factor non-normality has an ignorable impact on discrimination estimation both on general and specific factors, as well as location. However, the number of items per specific factor (*Is*) and sample size (*N*) have significantly influenced the result. This indicates that although non-normality may induce bias, increasing the number of items and the sample size can enhance the accuracy of item parameter estimations. The interaction between *Is* and *N* further emphasizes the significance of sample size concerns when addressing a fluctuating number of items per factor.

In our research, we used the two estimation algorithms, MAP and ML, to estimate the person parameter in Bifactor-GRM. The result exhibited that the MAP algorithm generally outperformed the ML algorithm in terms of both bias and RMSE for estimating general-factor and specific-factor abilities (*θg* and *θs*), especially for estimating specific-factor abilities (*θs*), if the number of items per specific factor is small. This suggests that MAP provides more reliable and accurate assessments, even with a limited number of items and a small sample size, while ML needs a sufficient number of items to get accurate person parameter estimates. Interestingly, while general-factor non-normality had a notable impact on the bias of a person’s general-factor ability estimates, its effect on the RMSE was relatively small. This suggests that while non-normality can shift the estimated values, it does not necessarily lead to a substantial increase in the variability of the estimates.

**Implications**

According to the results, there are several important implications for psychological and educational applications. To adopt the Bifactor-GRM, researchers and practitioners must acknowledge the skewness and kurtosis on both general- and specific-factor abilities possibly causing inaccuracy on item parameter recovery. Secondly, our findinNorms strongly advocate for the use of MAP estimation in the recovery of person parameters, particularly when the number of items per specific factor is limited, especially less than 10 items. The outstanding performance of the MAP algorithm for estimating person parameters indicates that it is a more robust method for addressing the issue caused by non-normality and limited data. Third, increasing the sample size and the number of items per specific factor could make the estimates of both item and person parameters more accurate, which would lessen the negative effects of non-normality.

**Limitations**

This study definitely has certain drawbacks which require additional research. During the simulation design phase, we exclusively simulated only one type of non-normality, namely skewness and kurtosis, and defined a fixed combination of their values. To enhance understanding of complex non-normality issues in real-world data, we may incorporate additional types of non-normality (e.g., bimodality) and utilize various combinations of skewness and kurtosis values. Furthermore, we exclusively examined constrained design factors of simulation study. Additional factors, like the number of factor loadings and the correlation between specific factors, could influence parameter recovery as well. Moreover, our research employed simulated data, which might limit the applicability of our results to real-world scenarios. Future research should examine the performance of the Bifactor-GRM in non-normal conditions using empirical data. Finally, we were able to evaluate the efficacy of psychometric software and R packages (e.g., FlexMirt, Mplus, and the R package lavaan) in Bifactor-GRM parameter recovery.

**References**

Auné, S. E., Abal, F. J. P., & Attorresi, H. F. (2020). A psychometric analysis from the Item Response Theory: step-by-step modelling of a Loneliness Scale. *Ciencias Psicológicas*, *14*(1).

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC press.

Benson, J., & Fleishman, J. A. (1994). The robustness of maximum likelihood and distribution-free estimators to non-normality in confirmatory factor analysis. *Quality and Quantity, 28*(2), 117-136.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in true data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78.

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18-27.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395-414.

Boulet, J. R. (1996). *The effect of nonnormal ability distributions on IRT parameter estimation using full-information and limited-information methods*. University of Ottawa (Canada).

Crișan, D. R., Tendeiro, J. N., Wanders, R. B., van Ravenzwaaij, D., Meijer, R. R., & Hartman, C. A. (2019). Practical consequences of model misfit when using rating scales to assess the severity of attention problems in children. *International journal of methods in psychiatric research*, *28*(4), e1795.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological methods*, *1*(1), 16.

DeMars, C. E. (2012). A comparison of limited-information and full-information methods in M plus for estimating item response theory parameters for nonnormal populations. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(4), 610-632.

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: a forgotten statistic. *Journal of statistics education*, *19*(2).

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and non-normality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 87-107.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521-532.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*(2), 91-115.

Heinz, A., Sischka, P. E., Catunda, C., Cosma, A., García-Moya, I., Lyyra, N., ... & Pickett, W. (2022). Item response theory and differential test functioning analysis of the HBSC-Symptom-Checklist across 46 countries. *BMC medical research methodology*, *22*(1), 1-24.

Hotelling, H., & Solomons, L. M. (1932). The limits of a measure of skewness. *The Annals of Mathematical Statistics*, *3*(2), 141-142.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. Structural Equation Modeling: A Multidisciplinary Journal, 5(4), 344-364.

Islam, M. Q., & Tiku, M. L. (2005). Multiple linear regression model under non-normality. *Communications in Statistics-Theory and Methods*, 33(10), 2443-2467.

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(1), 183-189.

Kehinde, O., Dai, S., & French, B. (2022). Item Parameter Estimations for Multidimensional Graded Response Model under Complex Structures. In *Frontiers in Education* (p. 597). Frontiers.

Lai, K. (2018). Estimating standardized SEM parameters given nonnormal data and incorrect model: Methods and comparison. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 600-620.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of non-normality in structural equation modeling. *Structural equation modeling*, *12*(1), 1-27.

Luh, W. M., & Guo, J. H. (2004). Improved robust test statistic based on trimmed means and Hall's transformation for two-way ANOVA models under non-normality. *Journal of Applied Statistics*, 31(6), 623-643.

Mardia, K. V. (1971). The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model. *Biometrika*, *58*(1), 105-121.

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal,* 24(3), 383-394.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, *105*(1), 156.

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of Bifactor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and Bifactor models via Monte Carlo simulations. *Journal of Intelligence*, *3*(1), 2-20.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and non-normality. *Structural equation modeling*, *7*(4), 557-595.

Ory, D. T., & Mokhtarian, P. L. (2010). The impact of non-normality, sample size and estimation technique on goodness-of-fit measures in structural equation modeling: evidence from ten empirical models of travel behavior. *Quality & Quantity*, 44, 427-445.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544-559.

Reise, S. P., & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. *Psychological Medicine, 46*(10), 2025-2039.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of personality assessment*, *98*(3), 223-237.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85-100). Springer, New York, NY.

Savalei, V. (2008). Is the ML chi-square ever robust to non-normality? A cautionary note with missing data. Structural Equation Modeling: A Multidisciplinary Journal, 15(1), 1-22.

Scherbaum, C. A., Cohen-Charash, Y., & Kern, M. J. (2006). Measuring general self-efficacy: A comparison of three measures using item response theory. *Educational and psychological measurement*, *66*(6), 1047-1063.

Seo, T., Kanda, T., & Fujikoshi, Y. (1995). The effects of non-normality of tests for dimensionality in canonical correlation and MANOVA models. *Journal of Multivariate Analysis*, 52(2), 325-337.

Sen, S., Cohen, A. S., & Kim, S. H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement*, 40(2), 98-113.

Sharma, K. K., Kumar, A., & Chaudhary, A. (2009). *Statistics in Management Studies*. Krishna Prakashan Media.

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and anxiety*, *25*(7), E34-E46.

Singh, A. K., Gewali, L. P., & Khatiwada, J. (2019). New measures of skewness of a probability distribution. *Open Journal of Statistics*, *9*(5), 601-621.

Svetina, D., Valdivia, A., Underhill, S., Dai, S., & Wang, X. (2017). Parameter recovery in multidimensional item response theory models under complexity and non-normality. *Applied psychological measurement*, *41*(7), 530-544.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological assessment*, *24*(1), 101.

Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of school psychology*, *60*, 41-63.

Urbán, R., Kun, B., Farkas, J., Paksi, B., Kökönyei, G., Unoka, Z., ... & Demetrovics, Z. (2014). Bifactor structural model of symptom checklists: SCL-90-R and Brief Symptom Inventory (BSI) in a non-clinical community sample. *Psychiatry research*, *216*(1), 146-154.

Van, den Oord, E. J., Pickles, A., & Waldman, I. D. (2003). Normal variation and abnormality: an empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry*, *44*(2), 180-192.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465-471.

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement*, *39*(8), 583-597.

Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate behavioral research*, *53*(3), 403-418.

White, H., & MacDonald, G. M. (1980). Some large-sample tests for non-normality in the linear regression model. *Journal of the American Statistical Association*, 75(369), 16-28.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological methods*, *11*(3), 253.

Xiao, Y., Liu, H., & Hau, K. T. (2019). A comparison of CFA, ESEM, and BSEM in test structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 665-677.