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

**Purpose of the Study**

In psychology and psychiatric research areas, it is common to encounter a latent construct that is positively skewed. For example, most people are in 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.

**Method**

**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,

**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 |
| GF Non-norm. (*GF*) | 2 | *GF =* Normal (Skew = 0, Kurt = 0),  Non-normal (Skew = 2, Kurt = 7) |
| SF Non-norm. (*SF*) | 2 | *SF =* Normal (Skew = 0, Kurt = 0),  Non-normal (Skew = 2, Kurt = 7) |

**Table 1Simulation Design**

|  |  |  |
| --- | --- | --- |
| Design Factors | Levels | Values of levels |
| **Data Structure** |  |  |
| Sample size (N) | 3 | N = 250, 500, 1000 |
| Number of Item per Factor (I) | 2 | I = 5, 10 |
| Number of Specific Factor (Fs) | 3 | Fs = 2, 3, 4 |
|  |  |  |
| **Nonnormality of Latent Traits (Theta)** |  |  |
| Nonnormality on general factor (θg) | 3 | Normal: Skewness = 0, Kurtosis = 0  Moderate: Skewness = 2, Kurtosis = 7  Severe: Skewness = 3, Kurtosis = 21 |
| Nonnormality on general factor (θsk) | 3 | Normal: Skewness = 0, Kurtosis = 0  Moderate: Skewness = 2, Kurtosis = 7  Severe: Skewness = 3, Kurtosis = 21 |

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 (*GF* = Normal, Non-normal) and the specific factor (*SF* = 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.

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 ability 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 parameter (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).

**Correlation**. Correlation measures the strength and direction of a linear relationship between the true personal ability and the estimated personal ability. A correlation closer to one indicates good performance of the estimation methods.

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**

For our analysis, we concentrated on those effects exceeding 0.06 to underscore the significant influences on item parameter estimation regarding bias and RMSE, as detailed in Table 2.

**Table 2**

*Analyzing Effect Sizes (Generalized Eta Squared> .06) of Main Effects and Interactions on Estimation Accuracy for Item Parameters*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Source of Variation |  | Bias |  |  | RMSE |  |
| *ag* | *as* | *c* | *ag* | *as* | *c* |
| *GF* | .498 | .001 | .167 | .024 | .000 | .010 |
| *Is* | .037 | .043 | .011 | .098 | .199 | .115 |
| *N* | .064 | .030 | .001 | .109 | .136 | .282 |
| *Is: N* | .036 | .039 | .000 | .055 | .082 | .054 |

***Note.*** The table displays effect sizes for bias and RMSE in item parameter estimation, which encompasses discrimination on the general factor (*ag*) and average discrimination on specific factors (*as*), as well as average location (*c*). The sources of variation encompass non-normality on general factor (*GF*), the number of items within specific factors (*Is*), sample size (*N*), and their interactions (*Is: N*).

In Table 2, the interaction between *Is* (the number of items within factors) ansd *N* (sample size) was the only interaction term that exhibited an effect size greater than 0.06. This interaction significantly influenced bias of estimation on *ag* (the discrimination on general factor), *as* (the average discrimination on specific factors), and RMSE of estimation on *ag*, *as*, and *c* (the average location), suggesting that *Is* impact estimation differently across various *N*. Regarding main effects, deviations from normality in the general factor primarily affected the bias (.498) in estimating *ag* and *c*, while having a less effect on the RMSE for all item parameter estimations. On the other hand, the number of items within factors significantly influenced the RMSE in estimating *as* and had a medium impact on estimating *c* but only minimally affected bias in estimating all item parameters. Additionally, *N* substantially affected the RMSE in estimating all item parameters and had medium effects on the bias in estimating *ag*.

**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 |  |  |  |  |  |  |
| *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) |
| *GF* | 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) |
| *RMSE* |  |  |  |  |  |  |
| *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) |
| *GF* | 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) |

***Note.*** The table displays means and standard deviations (SDs, in parentheses) of bias and RMSE under various conditions in item parameter estimation, which encompasses discrimination on the general factor (*ag*) and average discrimination on specific factors (*as*), as well as locations (*c1, c2, and c3*). The sources of variation encompass all conditions of general factor (*GF*), the number of items within specific factors (*Is*), sample size (*N*).

In Table 3, general factor discrimination (*ag*) shows generally small to moderate negative bias across all conditions, with non-normal distribution conditions on general factor showing the most significant bias (-0.19). Bias tends to become slightly more negative with larger sample sizes, reflecting a possible consistency in underestimation. Interestingly, for specific factor discrimination (*as*), bias remains very low across most conditions, with a slight positive bias at a smaller number of items (0.04 when *Is* = 5) and tending towards zero or slight negative bias as conditions change, indicating non-normality on general factor does not impact estimation on specific factor. Additionally, for the location parameters (*c1*, *c2*, and *c3*), bias is consistently negative across different conditions. Notably, under non-normal distributions of the general factor, these biases are more pronounced (*c1* = -0.10, *c2* = -0.10, and *c3* = -0.07), indicating a significant tendency to underestimate these parameters compared to other conditions. Generally, non-normal distributions on general factor and smaller sample sizes tend to exhibit slightly larger biases.

For RMSE, the general factor discrimination (*ag*) typically shows a decreasing trend as both the number of items per specific factor and sample size increase, highlighting improved estimation accuracy with larger data sets. Elevated RMSE values under non-normal conditions for the general factor suggest that such distributions lead to greater estimation errors. Similarly, RMSE for specific factor discrimination (*as*) decreases with an increase in the number of items and sample sizes, achieving its lowest values under conditions of large datasets and normal distribution for the general factor. For location parameters (c1, c2, and c3), RMSE generally declines with more items and larger samples, indicating that more extensive datasets yield more reliable estimations. Notably, RMSE for c1 shows a significant increase in non-normal conditions, contrasting with c3, where the impact appears less pronounced. This variation highlights the importance of robust sample characteristics and the normality of the distribution on the general factor in improving the precision of estimations.

**Figure 1**

*Interaction Plots of Bias of Item Parameters Estimation*

**A graph of different points

Description automatically generated with medium confidence**

*Note.*This figure demonstrates the bias in estimating item parameters across varying sample sizes for items 5, 10, and 20. It compares the discrimination on the general factor (*ag*), average discrimination on the specific factor (*as*), and the average threshold (*c*) under normal (solid line) and non-normal (dashed line) distributions.

In terms of estimation bias, the first row of Figure 1 highlights the effects on *αg*. It reveals that with normally distributed *θg* (theta, representing individual latent traits or abilities on the general factor), increasing the sample size results in a decrease in the estimation bias of *θg*, regardless of the number of items in a specific factor. Conversely, when *θg* are not normally distributed, larger sample size can exacerbate the bias (increase the absolute value) of *αg* across varying numbers of items in a specific factor. Additionally, the number of items in a specific factor can mitigate the bias of *αg* when *θg* are normally distributed. However, it tends to amplify the bias (increase the absolute value) when *θg* deviates from a non-normal distribution. Furthermore, except for a sample size of 250, non-normality of *θg* can lead to an increase in bias across the number of items in a specific factor.

The second row of Figure 2 demonstrates that the bias of estimating *αs* parameter is not significantly affected by the non-normality in the general factor. Only when specific factors consist of 5 items does sample size contribute to a reduction estimation bias of *αs*. When the specific factors have more than 10 items, the bias of discrimination on the specific factor approaches zero across all sample sizes, even with non-normality in the general factor.

In terms of c parameter (the average item location parameter), the last row of Figure 1 indicates that non-normality in the general factor can significantly influence the bias in estimating c parameter. However, when specific factors consist of 5 items, a small sample size of 250 can increase the bias (absolute value) of estimating *c*.

**Figure 2**

*Interaction Plots of RMSE of Item Parameters Estimation*

A graph of different locations

Description automatically generated with medium confidence

***Note.*** This figure demonstrates the RMSE in estimating item parameters across varying sample sizes for items 5, 10, and 20. It compares the discrimination on the general factor (*ag*), average discrimination on the specific factor (*as*), and the average threshold (*c*) under normal (solid line) and non-normal (dashed line) distributions.

Regarding RMSE, the first row of Figure 2 presents that the estimation RMSE of *αg* decreases with an increase in both sample size and item number in a specific factor. Normally distributed *θg*yields a greater reduction in RMSE compared to a non-normally distributed *θg*. Notably, when item number in specific factor exceeds 10 and *θg* is non-normally distributed, the RMSE of *αg* is only marginally affected by sample size.

The second row indicates that the estimation RMSE of *αs*decreases with an increase in sample size and item number in specific factor, regardless of whether *θg* is a normally or non-normally distributed. Only when there are 5 items in a specific factor is and the sample size is 250 does non-normality in *θg* significantly affect the estimation RMSE of *αs*.

The last row of Figure 2 addresses the impact of non-normality in the general factor on the bias in estimating the average item threshold parameter, *c*. Non-normality of *θg* does not significantly influence the bias in estimating *c*, but sample size and item number in a specific factor can.

**Person Parameter Estimation**

**Table 4**

*Effect sizes (> .06) of main effects and interactions for the bias of bias, RMSE and Correlation of estimation on person parameters*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Source of Variation | Bias | | | RMSE | | | | Correlation | | |
| *θg* | *θs* |  | | *θg* | *θs* |  | | *θg* | *θs* |
| *GF* | **.163** | .003 |  | | .039 | **.082** |  | | **.165** | .006 |
| *SF* | .001 | .047 |  | | .009 | .038 |  | | .006 | **.128** |
| *Fs* | .004 | .036 |  | | **.539** | **.657** |  | | **.681** | .005 |
| *Is* | .012 | **.092** |  | | **.742** | **.880** |  | | **.826** | **.836** |
| *N* | .002 | .000 |  | | .005 | .016 |  | | .016 | **.158** |
| *Method* | .042 | .020 |  | | **.856** | **.967** |  | | **.827** | **.360** |
| *SF: Is* | .000 | .004 |  | | **.117** | **.168** |  | | **.216** | **.066** |
| *Is: N* | .000 | .002 |  | | .041 | .014 |  | | .020 | **.123** |
| *Is: GF* | .004 | **.091** |  | | .002 | .005 |  | | .042 | .002 |
| *Fs: Method* | .000 | .028 |  | | **.264** | **.179** |  | | **.179** | .008 |
| *Is: Method* | .001 | **.093** |  | | **.589** | **.826** |  | | **.537** | .043 |
| *GF: Method* | .031 | .000 |  | | .006 | **.074** |  | | .022 | .001 |
| *Fs: N: GF* | .000 | .001 |  | | .001 | .009 |  | | .016 | **.071** |
| *Is: N: SF* | .004 | .002 |  | | **.072** | .011 |  | | **.072** | .035 |
| *Fs: Is: Method* | .001 | .004 |  | | **.079** | **.085** |  | | .020 | .005 |
| *Is: GF: Method* | .003 | **.093** |  | | .001 | .005 |  | | .003 | .001 |
| *Fs: Is: N: GF* | .003 | .003 |  | | .011 | .029 |  | | .032 | **.101** |
| *Fs: Is: N: GF: GS* | .001 | .001 |  | | .049 | .003 |  | | **.063** | .012 |

***Note.*** The table presents effect sizes for bias, RMSE, and correlation in the estimation of person parameters, including abilities associated with both general factors (*θg*) and specific factors (*θs*). The sources of variance considered include non-normality in general (*GF*) and specific factors (*SF*), the number of specific factors (*Fs*), the number of items within factors (*Is*), sample size (*N*), and their interactions. In the estimation of person parameters, two *Methods* (ML and MAP) are employed as within-subject factors in this ANOVA test. Non-normality refers to deviations from a normal distribution in abilities or traits. Interactions are examined to explore how item sequencing and sample characteristics influence estimation accuracy.

Table 4 reveals several critical insights in examining the influence of non-normality, specific factor number (*Fs*), the number of items in each specific factor (*Is*), sample size (*N*), and choice of algorithm on the estimation of item parameters. Notably, the algorithm used exerted the most substantial impact on both the RMSE for estimating general (*θg*) and specific (*θs*) factors, with RMSE values of .856 and .967, respectively, and a correlation for *θg* and *θs* estimation of .827 and .360, respectively. This suggests that the choice of algorithm is pivotal in achieving accurate item parameter estimations.

Furthermore, the interaction between the number of items in each specific factor (*Is*) and the algorithm significantly affected the RMSE for both *θg* and *θs* estimations, indicating that the complexity of model specifications can profoundly influence estimation accuracy. Specifically, the correlation between estimated *θg* and simulated *θg* for this interaction was notably high (.537), underscoring the need to simultaneously consider these factors to enhance the accuracy of the model.

Interestingly, the presence of non-normality in general factors (*GF*) and specific factors (*GF*) showed distinct effects on the bias and RMSE, highlighting the nuanced role that distributional characteristics play in psychometric analysis. Particularly, *GF* had a significant impact on the Bias of θg estimation (.163) and the RMSE of θs estimation (.082). Moreover, *GF* had a more pronounced effect on the correlation between estimated and simulated θg (.165), compared to *SF*, which significantly impacted the correlation for θs (.128).

The findings also illustrate the critical role of specific factor number (*Fs*), item number in each specific factor (*Is*) and sample size (*N*) influencing the RMSE and correlation outcomes, with larger effects observed for item number across the estimation accuracy metrics. This reinforces the notion that both the quantity of items (including *Fs* and *Is*) and the breadth of the sample are fundamental considerations in the construction and evaluation of psychological and educational assessments.

**Table 5**

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable and Condition | | Bias | | | RMSE | | | Correlation | |
| *θg* | *θs* | *θg* | | *θs* | *θg* | | *θs* |
| *Factor* | 2 | 0.03(0.06) | 0.00(0.08) | 0.61(0.38) | | 1.17(0.70) | 0.87(0.09) | | 0.48(0.16) |
|  | 4 | 0.02(0.06) | 0.02(0.07) | 0.42(0.23) | | 0.90(0.56) | 0.93(0.06) | | 0.47(0.13) |
| *Is* | 5 | 0.03(0.06) | -.02(0.09) | 0.71(0.42) | | 1.37(0.84) | 0.84(0.10) | | 0.33(0.10) |
|  | 10 | 0.02(0.06) | 0.01(0.06) | 0.48(0.23) | | 1.01(0.52) | 0.90(0.06) | | 0.48(0.08) |
|  | 20 | 0.02(0.06) | 0.02(0.05) | 0.35(0.14) | | 0.74(0.31) | 0.95(0.03) | | 0.62(0.08) |
| *N* | 250 | 0.02(0.07) | 0.01(0.08) | 0.51(0.30) | | 1.02(0.63) | 0.90(0.08) | | 0.45(0.17) |
|  | 500 | 0.02(0.06) | 0.01(0.07) | 0.52(0.34) | | 1.05(0.66) | 0.89(0.09) | | 0.47(0.15) |
|  | 1000 | 0.03(0.05) | 0.00(0.06) | 0.52(0.33) | | 1.04(0.66) | 0.90(0.07) | | 0.50(0.12) |
| *GF* | Normal | 0.00(0.06) | 0.01(0.06) | 0.50(0.32) | | 1.07(0.67) | 0.90(0.08) | | 0.47(0.15) |
|  | Non-N | 0.05(0.06) | 0.00(0.08) | 0.53(0.33) | | 1.01(0.63) | 0.89(0.08) | | 0.48(0.15) |
| *GS* | Normal | 0.03(0.06) | -.01(0.07) | 0.52(0.34) | | 1.02(0.65) | 0.90(0.08) | | 0.49(0.15) |
|  | Non-N | 0.02(0.06) | 0.02(0.07) | 0.51(0.32) | | 1.06(0.65) | 0.90(0.08) | | 0.45(0.14) |
| *Method* | MAP | 0.01(0.05) | 0.00(0.03) | 0.30(0.07) | | 0.52(0.11) | 0.94(0.04) | | 0.51(0.15) |
|  | ML | 0.04(0.06) | 0.01(0.10) | 0.73(0.34) | | 1.56(0.54) | 0.85(0.09) | | 0.43(0.13) |

***Note.*** The table displays means and standard deviations (SDs, in parentheses) of bias and RMSE under various conditions in person parameter estimation, which encompasses person traits/abilities on the general factor (*θg*) and the specific factor (*θs*), including normal (Normal) and non-normal (Non-N). The sources of variation encompass all conditions of the number of specific factors, number of items within specific factors (*Is*), sample size (*N*), normality on general factor (*GF*) and specific factor (*GS*), as well as *Method*, including Maximum A Posteriori (MAP) and Maximum Likelihood (ML).

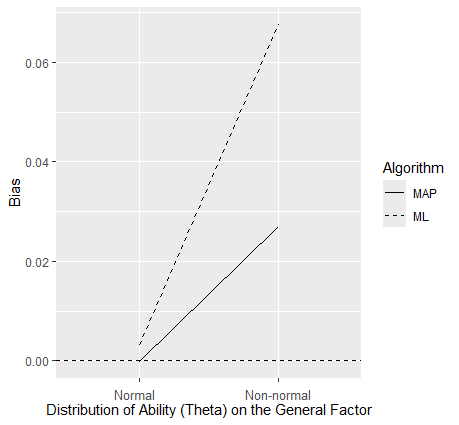
In Table 5, the person’s trait/ ability on general factor (*θg*) shows generally small to positive bias across all conditions, with non-normal distribution conditions on general factor showing the most significant bias (0.05), compared to normal distribution condition (0.00), and ML method exhibits higher bias (0.04) than MAP method (0.01). Meanwhile, the person’s traits/ abilities on specific factor (*θs*) are generally close to zero across all conditions, meaning estimation of *θs* is accurate without being affected by other factors.

For RMSE, the person trait/ability on the general factor (*θg*) consistently exhibited lower error rates compared to the specific factor (*θs*) across various conditions. Notably, RMSE for *θg* decreased with an increased number of items and larger sample sizes, reaching its lowest levels when the MAP method was employed, underscoring its effectiveness in providing precise estimates for the general factor. Conversely, RMSE values for the specific factor (*θs*) were consistently higher, suggesting greater variability and less precision in estimating these 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.

For correlation, *θg* are consistently high, indicating a strong and reliable match between the estimated and actual values of *θg*. This suggests that estimations of the general factor are robust across various conditions, including changes in the number of factors, items, sample sizes, and estimation methods. In contrast, the correlations for *θs* are generally lower than those for the general factor, reflecting more variability and reduced precision in the estimation of *θs*. However, the correlation for *θs* tends to improve significantly with increases in the number of items and in conditions utilizing the MAP method. This improvement highlights that while estimations for specific factors are less consistent than for the general factor, they can be enhanced substantially through methodological adjustments.

**Figure 3**

*Bias of Person Location (𝜃) Estimation on General and Specific Factor*

A graph with a line

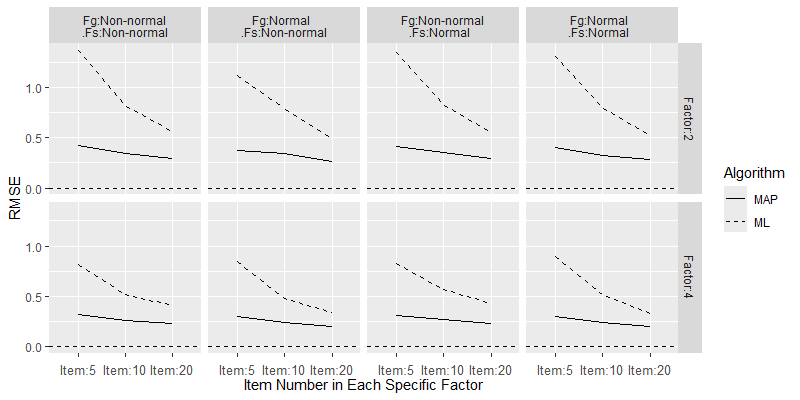
Description automatically generated

***Note.*** The Maximum A Posteriori (MAP) estimation method is depicted by the solid line, while the Maximum Likelihood (ML) estimation method is represented by the dashed line. The left figure relates to the general factor, and the right figure to the specific factor.

In Figure 3, on the left side, where a normal distribution of theta on the general factor is depicted, both MAP and ML algorithms start with equivalent biases at the origin point. However, as the distribution shifts from normal to non-normal, a clear divergence between the algorithms emerges. The ML estimation shows a pronounced increase in bias, illustrated by the steeper gradient of the dashed line, whereas the MAP estimation exhibits a more moderate rise in bias, as indicated by the solid line's less steep ascent. This visual contrast highlights the differential sensitivity of the MAP and ML estimations to deviations from a normal distribution in theta on general factor, revealing a complex interaction between the chosen algorithm and distribution characteristics that affects estimation bias in the assessment of person location. On the right side of Figure 3, while the MAP algorithm's performance remains consistent between normal and non-normal distributions of theta on the specific factor, the ML algorithm demonstrates a significant increase in bias, depicted by the steeply rising dashed line.

**Figure 4**

*RMSE of Person Location (𝜃) Estimation on General Factor*

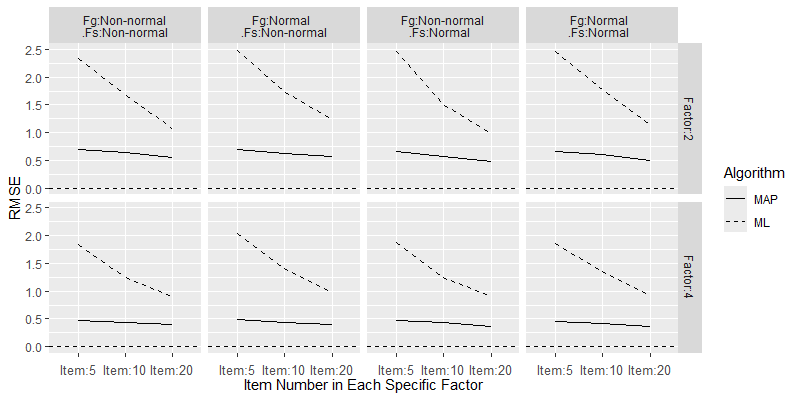


***Note.*** RMSE values are shown for different combinations of item numbers within specific factors (Item 5, Item 10, Item 20) across two conditions of factor distributions: normal and non-normal on the general factor (*Fg*) and specific factor (*Fs*). The analysis is further stratified by the number of specific factors (Factor 2 and Factor 4). Solid lines indicate MAP estimates, while dashed lines represent ML estimates.

Figure 4 shows that across all panels, the RMSE for both the MAP and ML algorithms generally decreases with an increasing number of items, indicating fewer estimation errors for higher item numbers. This trend is consistent across both levels of specific factors and all distribution conditions. The RMSE values are consistently higher for the ML algorithm than for the MAP for all item numbers and conditions, suggesting that the ML algorithm is more sensitive to the distributional characteristics of the items compared to the MAP algorithm. Additionally, the number of specific factors (Factor) appears to influence the RMSE, with four specific factors resulting in lower estimation errors than two specific factors, demonstrating that a greater number of specific factors can improve the accuracy of both MAP and ML algorithms.

**Figure 5**

*RMSE of Person Location (𝜃) Estimation on Specific Factor*



***Note.*** RMSE values are shown for different combinations of item numbers within specific factors (Item 5, Item 10, Item 20) across two conditions of factor distributions: normal and non-normal on the general factor (*Fg*) and specific factor (*Fs*). The analysis is further stratified by the number of specific factors (Factor 2 and Factor 4). Solid lines indicate MAP estimates, while dashed lines represent ML estimates.

Figure 5 illustrates the RMSE of person location (*θ*) estimation on specific factors. Similar to the findings in Figure 4, across all panels, the RMSE for both the MAP and ML algorithms tends to decrease as the number of items increases. This suggests that there are fewer estimation errors with higher item numbers, a pattern that holds true across both levels of specific factors and under all distribution conditions. The RMSE values for the ML algorithm are consistently higher than those for the MAP algorithm across all item numbers and conditions, implying that the ML algorithm may have a greater sensitivity to the distributional properties of the items compared to the MAP algorithm. Furthermore, the quantity of specific factors, termed Factor, seems to play a role in the RMSE; a setup with four specific factors yields lower estimation errors compared to one with only two specific factors. This indicates that an increase in the number of specific factors can enhance the estimation accuracy for both the MAP and ML algorithms.

**Figure 6**

*Correlation between Estimated and Simulation-Derived Person Location (𝜃) on the General Factor*A graph of different items

Description automatically generated with medium confidence

***Note.*** Correlation of *𝜃* on the general factor are shown for different combinations of item numbers within specific factors (Item 5, Item 10, Item 20) across two conditions of factor distributions: normal and non-normal on the general factor (*Fg*) and specific factor (*Fs*). The analysis is further stratified by the number of specific factors (Factor 2 and Factor 4). Solid lines indicate MAP estimates, while dashed lines represent ML estimates.

In Figure 6, all panels show a pattern where correlation generally improves (increases toward 1, indicating perfect correlation) with an increasing number of items. The MAP algorithm consistently demonstrates higher correlation values than the ML algorithm, suggesting that MAP estimates are more closely aligned with the simulation-derived values. Variations between normal and non-normal distributions of the general and specific factors reveal that non-normal distributions tend to have a lower correlation, especially on the general factor, indicating a reduced alignment between estimated and simulated person locations under these conditions. The comparison between the two algorithms suggests that MAP maintains a more robust correlation across various conditions and number of items in each specific factor, indicating its potential superiority over ML in maintaining estimation accuracy in the context of the general factor. Additionally, the number of specific factors also plays a role in the RMSE: a setup with four specific factors yields higher correlation compared to one with only two specific factors. This indicates that an increase in the number of specific factors can enhance the estimation accuracy for both the MAP and ML algorithms.

**Figure 7**

*Correlation between Estimated and Simulation-Derived Person Location (𝜃) on the Specific Factor*

A graph of different items

Description automatically generated with medium confidence

***Note.*** Correlation of *𝜃* on the specific factor are shown for different combinations of item numbers within specific factors (Item 5, Item 10, Item 20) across two conditions of factor distributions: normal and non-normal on the general factor (*Fg*) and specific factor (*Fs*). The analysis is further stratified by the number of specific factors (Factor 2 and Factor 4). Solid lines indicate MAP estimates, while dashed lines represent ML estimates.

Figure 7 illustrates the correlation between estimated and simulation-derived person location (θ) on the specific factor. As observed in the panels, there is a trend where the correlation typically improves with an increasing number of items, approaching 1. The MAP algorithm consistently shows higher correlation values compared to the ML algorithm across the board, implying that estimates from MAP are more congruent with the values derived from simulation. Considering the distribution conditions of the specific factor, the panels indicate that non-normal and normal distributions do not yield a pronounced difference on correlation. Moreover, similar to the trend in general factor estimations, the number of specific factors affects the correlation: setups with four specific factors demonstrate a higher correlation than those with only two specific factors, reinforcing the idea that an increased number of specific factors can lead to improved estimation precision for both MAP and ML algorithms.

**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.