**Impact of Latent Traits Non-normality on Parameter Estimation in Bifactor Graded Response Model**

**Abstract**

Psychometric models, such as item response theory (IRT), usually assume a normal distribution of latent abilities or traits. Nonetheless, violations of normality assumption frequently occur in psychology and psychiatric research. Many researchers have investigated the impact of non-normality on IRT and other psychometric models, but insufficient attention has been given to its influence on bifactor IRT models, particularly with polytomous data. In this study, we address the problem of non-normality in parameter estimation within the framework of the bifactor-graded response model (Bifactor-GRM). The results indicate that non-normality in the general factor significantly impacts the accuracy of estimating general-factor discrimination and thresholds/locations but not limited impacts on specific-factor parameters. If specific factors have the non-normality issue, there are minor impacts on item parameter estimation for both the general and specific factors. Additionally, we found that increasing sample sizes and item numbers may reduce the impact of non-normality. Regarding person parameter estimation, the maximum a posteriori (MAP) algorithm generally outperforms marginal maximum likelihood (ML), particularly when population abilities or traits are non-normally distributed. To enhance the accuracy of the ML algorithm, researchers could increase sample sizes and item numbers.

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

**Introduction**

In psychological and educational measurement, the bifactor item response theory (IRT) model (Gibbons & Hedeker, 1992) has gained considerable attention in the past decades because of its ability to simultaneously model a general factor and specific factors underlying multidimensional constructs (DeMars, 2013; Heinrich at al., 2023; Reise at al., 2023; Rodriguez at al. 2016). In a bifactor model, item responses are a function of a general factor and a specific factor, and the specific factors are orthogonal to the general factor. Bifactor models have been widely applied in areas such as cognitive ability testing, personality assessment, and clinical psychology, where a separation between primary and secondary factors is preferred (add some citations here).

Many IRT models including bifactor models are estimated under the assumption that the latent traits conform to a (multivariate) normal distribution. This assumption is fundamental to the widely used marginal maximum likelihood estimation method (Bock & Aitkin, 1981; Bock & Lieberman, 1970), which treats latent trait as a random variable rather than a fixed parameter. The (multivariate) normal distribution is mathematically convenient and often serves as a reasonable approximation for many latent traits, particularly in the context of educational assessment. However, in many psychological and psychiatric contexts, this assumption may not be valid. For instance, in clinical assessments of psychiatric disorders, the majority of individuals tend to cluster at the lower end of the severity spectrum, while only a small subset exhibits extreme symptoms, leading to a positively skewed latent trait distribution. Similar patterns emerge in assessments of rare cognitive deficits or personality traits, where most individuals score within a typical range, and only a few deviate substantially. When the assumption of normality is violated, the accuracy of parameter estimation in IRT models may be compromised.

The impact of the non-normality of latent traits on parameter estimation has attracted attentions (e.g., Wang et al., 2018) in the field of structural equation modeling (Finch et al., 1997; Lai, 2018; Lei & Lomax, 2005; Maydeu-Olivares, 2017; Olsson et al., 2000; Ory & Mokhtarian, 2010) and factor analysis (Curran et al., 1996; Hutchinson & Olmos, 1998; Savalei, 2008), as well as IRT (Svetina et al., 2017; Woods, 2014). Researchers have found that., in IRT models, non-normality in the latent trait distribution can significantly decrease the accuracy of item and person parameter recovery. In the Maximum Likelihood Estimation (MLE) method, non-normality tends to produce biased estimates, especially at the tails of the distribution (Kieftenbeld & Natesan, 2012; Selçuk & Demir, 2024; Seong, 1990). In contrast, in the Bayesian framework, Maximum a Posteriori (MAP) and Expected a Posteriori (EAP) methods can partially mitigate non-normality via incorporating the sample’s prior distribution, but their performance is still affected when the sample size is insufficient to detect non-normality (Finch & Edwards, 2016; Kieftenbeld & Natesan, 2012; Selçuk & Demir, 2024). Additionally, Sass et al. (2008) addressed the fact that non-normal distributions require attention to item difficulty, discrimination estimates, and person latent trait estimates across all methods.

Although a few studies have examined the impact of non-normality of latent trait under unidimensional and multidimensional IRT models, the impact on estimating parameters of bifactor IRT model remains unstudied. Bifactor models are unique in terms of their orthogonal structure of general and specific factors. When the latent distributions are non-normal, the orthogonality assumption may no longer be appropriate or realistic. For instance, a skewed general trait and symmetric specifics can confound the source of variance. Bonifay and his colleagues (2017) noted that bifactor models are highly sensitive to data features like skewness or limited variability. Moreover, bifactor models estimate general and specific latent attributes simultaneously, assuming normal latent trait distribution for both factors. The violation of the normality assumption from either general or specific factors possibly impacts the estimation of both factors.

Previous bifactor IRT simulation studies generally applied non-normality (e.g., skewness or kurtosis) exclusively to the general-factor latent traits, neglecting specific-factor traits (e.g., Wang et al., 2018). However, the non-normality of specific factors may also affect item responses. The simultaneous estimation of latent traits in both general and specific factors can lead to misleading parameter estimates because of violations of normality in the specific factors (Bonifay, Lane, & Reise, 2017). Furthermore, the interaction effects between general non-normality and specific variables are inadequately examined in the present research, leaving a significant gap in understanding the influence of non-normality on estimating bifactor IRT models.

In addition, previous research found that Maximum a Posteriori (MAP) estimation has demonstrated superior performance over Maximum Likelihood (ML) under non-normality conditions, producing lower bias and root mean square error (RMSE) in both item and person parameter estimates (Wang et al., 2018; Kieftenbeld & Natesan, 2012). However, Bayesian estimation methods (e.g., MAP and EAP) rely on informative priors; small sample sizes may not sufficiently adjust for skewed latent attribute distributions, while each specific factor, with even smaller items, could face more severe impacts.

The goals of this paper are twofold: first, to examine the effects of violating the normality assumption on parameter estimation in bifactor IRT models; second, to evaluate the effectiveness of MAP and ML methods in estimating latent traits of individuals across different simulated data conditions, such as sample size, item number and number of specific factors. Specifically, this simulation study examines how non-normality in both general and specific factors impacts the recovery of item and person parameters. It also considers the often-neglected effects of non-normality in specific factors and investigates the interaction effects between non-normality in general and specific factors. This study is focused on clarifying the robustness and application of the MAP method in bifactor-GRM modelling while investigating the effects of sample size, item number, number of specific factors, and distribution conditions on both general and specific factors.

This paper is structured as follows. Initially, we introduce the bifactor graded response model (Bifactor-GRM) within the framework of item response theory. The following part defines the simulation design, including the manipulated variables and their respective conditions. The results section is separated into two parts: the item parameter estimate and person parameter estimation. Item parameters estimation, including bias and root mean square error (RMSE), are presented to assess the influence of non-normality on both general and specific aspects. Bias and RMSE are also used to assess the performance differences between the MAP and ML estimation methods for person parameters.

**Bifactor Grade Response Model**

The Bifactor-GRM (Gibbons et al., 2007; Reise et al., 2010) is an extension of the conventional GRM (Samejima, 1969) and is suitable for handling ordinal response data. Suppose a test involves one general factor and S specific factors and each item is related to the general factor and one of the S specific factors (Toland et al., 2017). Assume item *j* is associated with a general factor *g* and a specific factor *s*, and responses to item *j* is polytomous with ordered categorical values of 1, …, *K*. The cumulative probability that a respondent *i*’s response falls at or above a particular ordered category, given latent trait *θ*, can be formulated as:

 (1)

in which, *θg* represents the respondents’ latent trait on the general factor, while *θs* denotes their latent trait on the specific factor. The parameters *ajg* and *ajs* are the discrimination of item *j* on the general and specific factors, respectively, and *bjk* is the location for category *k*. The constant *D* is added as a scaling factor, typically set to 1 or 1.7.

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,

 (2)

in which,  and are two adjacent cumulative probabilities.

**Simulation Study**

**Design Factors**

This study conducts a Monte Carlo simulation to assess the impacts of non-normality issue on both factor and specific factor within a bifactor model. Additionally, the research incorporates four-category polytomous items, which are frequently used in psychometric and educational assessments. Drawing on methodologies established in prior studies (Auné, 2020; Mao, 2022; Rijmen, 2011; Svetina et al., 2017; Wang et al., 2018), this study will summarize common and practical non-normality issues and explore how deviations from normality can affect bifactor model parameters estimation To observe this, a total of five variables were manipulated, as shown in Table 1.

**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 |
| Normality condition of the general factor (*Normg*) | 2 | Normal : Skew = 0, Kurt = 0,  Non-normal : Skew = 2, Kurt = 7 |
| Normality condition of the specific factor (*Norms*) | 2 | Normal : Skew = 0, Kurt = 0,  Non-normal : Skew = 2, Kurt = 7 |

The sample size has three levels (*N* = 250, 500, 1000), the number of items per factor has three levels (*Is* = 5, 10, 20), and specific factor has two levels (*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 commonly adopted in previous research (Curran et al., 1996; Wang et al., 2018). For each condition, we set all specific factors fixed at same level of skewness and kurtosis.

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 parameters are usually positive and fall within the range of 1.1 to 2.8 (Atkinson, 2018; Auné, 2020; Rijmen,2011). Previous studies have consistently shown that specific factor discrimination parameters are typically smaller than those of the general factor, generally ranging from 0 to 1.5 (Wang et al., 2018). Given the assumption of orthogonality of the latent traits on the general factor and specific factor, we generated two sets of latent trait for each factor, separately. The discrimination parameter for the general factor is randomly set to range from 1.1 to 2.8, while the discrimination parameter for the specific factors is kept within the range of 0 to 1.5.

Although item difficulty values can theoretically range from negative infinity to positive infinity, they practically locate from -2 to +2 (Hambleton & Swaminathan, 1985). Psychological and psychiatric assessments often use a four-point Likert scale to measure latent traits or personalities (Auné et al., 2020; Rijmen,2011). Building on the setting in Wang’s (2018) prior research, this study generated normally distributed item locations following three intervals: b1[−2, −0.67], b2[−0.67, 0.67], and b3[0.67, 2], to distinguish the possibilities of choosing each response category.

**Person parameter**

The values for skewness between -2 and +2 are considered acceptable for assuming normality (Hair et al., 2010). Additionally, kurtosis is another critical indicator of distribution sharp, and values located between -7 to +7 are generally considered as normality (Hair et al., 2010;). Therefore, 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 conditions across the general factor and specific factor, reflecting whether each was normal or non-normal respectively. In this study, we employed the Fleishman method to generate nonnormal distributions, involving 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". Due to complex and non-normal data structures, the study increase the iteration limit to 6000 to ensure accurate parameter recovery. 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 on person ability parameters involved utilizing the "fscores()" function. It is important to note that, in “mirt” package, the location parameters are outputted as *djk* such that *bjk* = *-djk*, where *bjk* refers to the location used in 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 item and person parameters. The item parameters include two discrimination parameters for general factor and specific factor as well as three location parameters. The person parameters consist of two latent traits corresponding to general and specific factors.

**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,

 (3)

where is the estimated parameters () across valid replications and is the true parameters (). In the and , *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 1000 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,

 (4)

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 small effect size of 0.01, moderate effect size of 0.06, and large effect size of 0.14 (Gignac & Szodorai, 2016).

**Results**

**Item Parameter Estimation**

In this section, we only looked at ANOVA effect size (*η²*) exceeding 0.06, to get a clear picture 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*: Normality condition of the general factor; *Is*: Item number per specific factor; *N*: Sample size; *Is× N*: Interaction between *Is* and *N*;*ag*: discrimination on the general factor; *as*: average discrimination of specific factors; *c*: average value of locations (*c1*, *c2*, and *c3*).

According to Table 2, non-normality in the general factor (*Normg = non-normal*) primarily causes significant bias in the estimation of discrimination on general factor (*ag*) and 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 location, significantly influencing RMSE of estimating the discrimination on specific factors (*as*). We will apply a plot to explain this interaction effect in the following section. In this table, there is no impact of non-normality in the specific factor (*Norms = non-normal*; *η²* < 0.06), meaning that non-normality on specific factor has only negligible effect on item parameter estimation.

**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*: Normality condition of the general factor; *Is*: Number of Item on specific factors; *N*: Sample size;*ag*: discrimination on general factor; *as*: average discrimination on specific factors; *c*: average value of locations (*c1*, *c2*, and *c3*).

Table 3 presents all means and standard deviations (SDs) of bias and RMSE in estimating all five item parameters (*ag*, *as*, *c1*, *c2*, *c3*) across all simulation conditions. Regarding bias, non-normality in the general factor (*Normg = non-normal*) led to noticeably increased negative bias in estimating general-factor discrimination (*ag*​), with a mean bias of –0.19 compared to 0.03 under normality. However, the bias for specific-factor discrimination (*as​*) was minimal across both normal and non-normal conditions. The bias of the locations (*c1*, *c2*, and *c3*) are all increased by presence of non-normality in the general factor. However, in this simulation study, non-normality in the general factor (*Normg = non-normal*) does not have a big impact on RMSE of estimation for any of the five item parameters (*ag*, *as*, *c1*, *c2*, *c3*). This is finding addressed that non-normality in the general factor can introduce systematic bias, but not change the variability of estimates, resulting in minimal impact on RMSE.

Additionally, as item number per specific factor (*Is*) go from 5 to 20, both bias and RMSE of estimation decrease, which suggests that having more items makes estimation more stable. Similarly, a sample size (*N*), increasing from 250 to 1000, also lowers bias and RMSE of estimation on all item parameters. Overall, the results suggest that non-normally distributed ability on the general factor, smaller item sets, and sample sizes all contribute to a 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*: average discrimination on specific factors; RMSE: Root Mean Square Error; Normality in general factor is depicted by the solid line, while non-normality in general factor is depicted by the dashed line.

Figure 1 shows that the RMSE for estimating the specific-factor discrimination parameter (*αs*) decreases as both the sample size and the number of items per specific factor increase, indicating both sample size and item number significantly improve the accuracy of estimation of *αs*. However, the existence of non-normality in general factor (dashed line) had little impact on RMSE of estimating specific-factor discrimination (*αs*). c

**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** |
| *Norms* | .001 | .047 |  | .009 | .038 |
| *Fs* | .004 | .036 |  | **.539** | **.657** |
| *Is* | .012 | **.092** |  | **.742** | **.880** |
| *N* | .002 | .000 |  | .005 | .016 |
| *Method* | .042 | .020 |  | **.856** | **.967** |
| *Norms × 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× Norms* | .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*: Normality condition of the general factor; *Norms*: Normality condition of the specific factor; *Fs*: Number of specific factors; *Is*: item number per specific factor ; *N*: Sample size; *Method*: Person parameter estimation algorithm (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 item number per specific factor (*Is*), sample size (*N*), and the algorithm for person parameter estimation (*Method*). In terms of bias, *Normg* accounted for the exclusive impact on *θg* estimation (*η2* = .163), while *Is* is the single factor which affects *θs* estimation (*η2* = .092). Regarding RMSE, the estimation algorithm had the largest effect on RMSE of both *θg* and *θs*, with *η2* values of .856 and .967, respectively. Following, *Is* impacts RMSE of both *θg* and *θs* with *η2* values of .742 and .880, and *Fs* with *η2* values of .539 and .657. Additionally, *Normg* cause minor but detectable RMSE on estimation of *θs*. This emphasizes the critical importance of selecting the MAP algorithm to improve the accuracy of person parameter recovery and illustrates how non-normality in the general factor reduces the accuracy of estimating latent traits for both general and specific factors.

Table 4 also reveals interaction effects, we will focus only two-way interactions here and leave three-way interactions analyzed with a graphical way in the next section. Regarding bias, only two two-way interactions are found to influence estimation of *θs* the: interaction between *Is* and *Normg*, as well as the interaction between *Is* and *Method*. Regarding RMSE, several two-way interactions account for variance in the estimation of both *θg* and *θs*. For *θg*, three two-way interactions were found: the interaction *Norms × Is* (*η2* = .117), *Fs× Method* (*η2* = .264), and *Is× Method* (*η2* = .589). Meanwhile, for *θs*, four two-way interactions are exhibited: the interaction *Norms × Is* (*η2* = .168), *Fs× Method* (*η2* = .179), *Is× Method* (*η2* = .826), and *Normg× Method* (*η2* = .074). The main effect of *Norms* is not statistically significant as a single factor, butit is a key component in several significant two-way interaction effects.

These findings underscore the importance of carefully selecting estimation algorithm (e.g., MAP), the number of specific-factor and item number per specific factors to enhance person parameter recovery, particularly when the Bifactor-GRM involves non-normality in either the general or specific factor.

**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.*** SD **=** Standard deviations (in parentheses). *Normg*: Normality condition of the general factor; *Norms*: Normality condition of the specific factor; *Fs*: Number of specific factors; *Is*: Item number 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 ability (*θg*) shows small 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). Another notable difference is that ML method exhibits higher bias (0.04) than MAP method (0.01), indicating the MAP method has better performance than ML on estimating *θg*. Meanwhile, the biases of persons’ specific-factor abilities (*θs*) are generally close to zero across all conditions, meaning that estimated *θs* are closed to real *θs* or symmetrically distributed around real *θs*.

Regarding RMSE, the *θg* consistently exhibited lower RMSE compared to *θs* across all conditions. Also, the non-normality in the general factor (*Normg* = Non-normal) affected estimation on *θg* slightly, increasing only marginally from normal distribution (0.50) to non-normal distribution (0.53). *Normg* did not increase RMSE of estimation on *θs*; in fact, it slightly decreased from 1.07 to 1.01. Conversely, the non-normality issue on the specific factor (*Norms* = Non-normal) increased RMSE of estimation on *θs* from 1.02 to 1.06, while slightly decreasing RMSE of *θg* from 0.52 to 0.51. Additionally, higher *Fs* and *Is* are significantly decrease the RMSE of estimation on both *θg* and *θs*. The most important factor is method, the MAP method has smaller RMSE on both *θg* (0.30) and *θs* (0.52), compared to the ML method for estimating *θg* (0.73) and *θs* (1.56). This result indicates that the MAP method is a better performance on ability estimation.

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

Figure 2 exhibits that the MAP method shows minor bias of estimating *𝜃g* with no significance between normality (solid line) and non-normality (dashed line) in the general factor. In contrast, although 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 in general factor.

Overall, the results demonstrate that MAP provides more stable estimates and is less affected by violations of normality assumption in the general factor, whereas ML produces significant biased estimates due to the presence of non-normality. Specifically, a smaller numbers of items per specific factor leads to notable negative bias, while a larger number 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 in general factor is depicted by the solid line, while non-normality in 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 *Is* had a noticeable effect on RMSE of person general-factor ability (*𝜃g*) estimation. This means that greater item number per specific factor can significantly reduce the negative impact of non-normality in the general factor.on 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-factor condition is depicted by the solid line, while the four-specific-factor condition is depicted by the dashed line.

Figure 4 compares the RMSE in general-factor ability (*𝜃g*) estimation across methods, item number per specific factor (*Is*), and specific factor number (*Fs*). The result shows that the MAP algorithm generally has lower and more consistent RMSE values across all conditions of item numbers, outperforming the ML method. Under the ML method, raising *Fs* and *Is* can effectively reduce the RMSE; however, this impact is slight under the MAP method. In conclusion, increasing the number of items per specific factor and/or specific factor number can ,to some extent, reduce the disparity in estimation performance between the MAP and ML approaches.

**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 number per specific factor (*Is*), and specific factor numbers (*Fs*). Similar to the RMSE in *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 *𝜃s* estimation is more sensitive to a small number of items within each specific factor.

**Discussions**

The non-normality in latent trait distributions is a common issue in psychological and educational assessment as forementioned. As a widely used model, the -GRM might face a more complicated estimation when violations of normality assumption can happen on both the general and the specific factor. This section will address the impact of violations of normality assumption in both item and person parameter estimation within the Bifactor-GRM and offer implications for both researchers and practitioners in psychometrical research or operational areas.

Our results revealed that for estimating item parameters of Bifactor-GRM, non-normality in the general factor (*Normg*) 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*). Regarding RMSE, *Normg* has an ignorable impact on the RMSE in estimating discrimination on general, specific factor, and location parameters. Interestingly, increasing the number of items per specific factor (*Is*) and sample size (*N*) can some extent offset the negative effect of non-normality. The only interaction found in item parameter estimation was between *Is* and *N* for RMSE, which disclosed that the number of item per specific factor and sample size play a critical role in enhancing the estimation on discrimination parameter on the specific factor (*as*), rather than on the general factor (*ag*).

For person parameter estimation, we used the MAP and ML computational algorithms. 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 keep accuracy of 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**

This simulation study addresses several practical implications for applying the Bifactor-GRM for psychological and educational assessments when the latent traits of the population have a non-normal distribution issue. First, researchers need to be mindful that non-normal latent trait distributions, specifically skewed or leptokurtic general and specific factors, can introduce systematic bias in item parameter recovery. This result agrees with prior studies, confirming that violation of the normality assumption can compromise the validity of parameter estimates in IRT and SEM frameworks (Curran et al., 1996; Wang et al., 2018). Therefore, when practitioners employ assessments in clinical or low-prevalence populations and notice that latent traits are skewed in preliminary analysis (e.g., psychiatric symptom severity), they should be cautious to interpret item parameters without considering the impact of the violation of the non-normality assumption.

Second, the findings strongly recommend the use of the MAP estimation method over traditional ML, especially when the number of items per specific factor is limited (fewer than 10). This conclusion is consistent with the conclusion from Finch and Edwards (2016), who found that Bayesian estimation methods are more robust under small sample sizes and violations of normality assumptions. Practically, if a profession decides to adopt short-form tests (e.g., subdomains with limited items), he could consider using the MAP method to achieve more reliable person score estimation.

Finally, in Bifactor-GRM, enlarging the sample size and number of items per specific factor generally enhanced the estimation accuracy for both item and person parameters. This finding corresponds with suggestions from earlier psychometric research that advise enough test length and respondent numbers to allow stable estimates, especially in multidimensional models (DeMars, 2012; Reise et al., 2010). These results underscore that the estimation accuracy of Bifactor-GRM is influenced not only by the statistical method employed but also by rigorous test design and assumption of normality in both general and individual factors.

**Limitations**

This study definitely has certain limitations that require additional research. During the simulation design phase, we exclusively simulated only one type of non-normality, 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 (e.g., skewness > 2 and kurtosis > 7). 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**

Atkinson, P. (2018). *The Clinical Experience, (1997): The Construction and Reconstrucion of Medical Reality*. Routledge.

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).

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179-197.

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184-186.

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.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International journal of testing, 13*(4), 354-378.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4-19.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and individual differences*, *102*, 74-78.

Fialkowski, A. C. (2018). SimMultiCorrData: Simulation of correlated data with multiple variable types. *R package version 0.2*, *2*(10.1002).

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.

Finch, H., & Edwards, J. M. (2016). Rasch model parameter estimation in the presence of a nonnormal latent trait using a nonparametric Bayesian approach. *Educational and Psychological Measurement*, *76*(4), 662-684. https://doi.org/10.1177/0013164415608418

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

Hambleton, R. K., & Swaminathan, H. (1985). Estimation of item and ability parameters. In *Item response theory: Principles and applications* (pp. 125-150). Dordrecht: Springer Netherlands.

Heinrich, M., Geiser, C., Zagorscak, P., Burns, G. L., Bohn, J., Becker, S. P., ... & Knaevelsrud, C. (2023). On the meaning of the “P factor” in symmetrical bifactor models of psychopathology: Recommendations for future research from the bifactor-(S− 1) perspective. Assessment, 30(3), 487-507.

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.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*(5), 399-419.

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.

Mao, X., Zhang, J., & Xin, T. (2022). The optimal design of bifactor multidimensional computerized adaptive testing with mixed-format items. *Applied Psychological Measurement*, *46*(7), 605-621.

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.

Reise, S. P., Mansolf, M., & Haviland, M. G. (2023). Bifactor measurement models. *Handbook of structural equation modeling*, 329-348.

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.

Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI monograph series: Issues and methodologies in large-scale assessments*, *4*, 59-74.

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.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological methods, 21*(2), 137.

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

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, *21*(1), 65-88.

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.

Selçuk, E., & Demir, E. (2024). Comparison of item response theory ability and item parameters according to classical and Bayesian estimation methods. *International Journal of Assessment Tools in Education*, *11*(2), 213-248.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied psychological measurement*, *14*(3), 299-311.

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.

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.

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.

Woods, C. M. (2014). Estimating the latent density in unidimensional IRT to permit non-normality. In *Handbook of item response theory modeling* (pp. 60-84). Routledge.