**Examine the Impact of Nonnormality on Parameter Estimates in the Multidimensional (Bi-factor) graded response model**

Jujia Li

Department of Educational Studies in Psychology,

Research Methodology, and Counseling

BER 699: Item Response Theory

Dr. Wencao Ma

December 8, 2022

**Abstract**

In psychological and psychometric research, the non-normality of the distribution of latent traits (θ) is a prevalent phenomenon. Most commercial software and open-source package offer one or more estimation methods to estimate the parameters of models, but most of them are based on the normal distribution. This study is simulation research focused on the impact of nonnormality on parameter estimates in the Bifactor IRT model in flexMIRT, involving 162 conditions (30 replications per condition). It can be concluded that the skewness of latent traits will affect the precision of estimation in flexMIRT, the skewness of latent traits in specific factors (θs) is more difficult to be detected in the model estimation, compared to general factor (θg). The bias, RMSE, and variance of each parameter is compared across different conditions via a repeated-measures ANOVA. Overall, sample size, and Item per factor, and skewness on general factor are third major factors.

**Keywords: Nonnormality, Bifactor, IRT**

**Introduction**

**Nonnormality of the multidimensional model**

In psychology and psychiatric research area, the impact of the nonnormality of the multidimensional IRT model is a prevalent phenomenon, attracting researchers’ attention and worry. Due to easy access, community-based sampling involves non-clinical samples in which the level of distress and the number of symptoms is lower than in a clinical sample are always used in previous research (Urbán, Kun, Farkas, Paksi, Kökönyei, Unoka... & Demetrovics, 2014). When adding on clinical samples, positively skewed models are typically in a general population, such as psychiatric disorder traits (Wall, Park, & Moustaki, 2015; Woods&Thissen, 2006). Specially, scale scores in studies of emotional and behavioral problems often possess highly skewed distributions (Oord, Pickles, & Waldman, 2003).

An analysis of several hundred psychometric and achievement data distributions in education and psychology found that 49% had at least one extremely heavy tail (Micceri, 1989). In another analysis of 693 distributions of cognitive ability and of other psychological variables, only 5.5% of distributions were close to expected values under normality (Blanca, María J. et al., 2013). In psychiatric clinical studies, including drug abuse research, data are always heavily skewed due to subjects’ self-reports, the need to assess illegal behaviors, and high rates of participant attrition (Delucchi & Bostrom, 2004).

Although transformations (e.g.logarithmic, Box-Cox) can make the data conform to the methods’ assumptions, it is not always possible to find a suitable one, and analyzing data on a different scale might compromise interpretability (Counsell, Cortina-Borja, Lehtonen, & Stein, 2010).

**IRT model**

Most psychometric tests’ questions are multiple-option, such as patient reported outcome (PRO) and health-related quality of life (HRQOL) research (Atkinson, et al., 2018), the Symptom Checklist (SCL, Heinz, et al., 2022), and UCLA Loneliness Scale (Auné, et al., 2020). When analyzing multiple-level scale, item response model is a prevalent method applied to analysis participants’ latent traits.

Most IRT makes no formal normality assumption, item parameters are typically estimated assuming a normal prior for the latent variable. Violating this assumption can only lead to biased parameter estimates (Reise & Rodriguez, 2016).

**Bifactor model**

Unidimensional models limit researchers to understand mental disease and find effective treatments. To make progress in reducing stigma against the mentally ill and improving the effectiveness of treatments, we possibly can seek to understand more multidimensional models and consider how multiple systems interact in contributing to mental disorders. Over the past decade, bifactor model analysis has become increasingly popular as a statistical approach to describe common and specific elements in psychology and psychiatry. Compared with unidimensional (61%) and oblique simple structure (65%) models, the bifactor model demonstrated a superior fit for the internal structure of the BSI and improved overall diagnostic accuracy in the sample (73%) (Thomas, 2012).

More than unidimensional structure models, the bifactor model performs better than other multidimensional models, such as high-level model and hierarchical model. Based on 100 published candidate bifactor articles, the authors reached many conclusions regarding their estimated bifactor models, but no one reported a “bad” or “poor” fit (Rodriguez, Reise, & Haviland, 2016). However, in recent years, some researchers addressed some drawbacks of the bifactor model, such as overfitting, small factor loadings, zero or negative group factor variances, instability of the general factor, problems with identification, interpretation of orthogonal latent factors, and reification in searches for genetic or biological substrates of the p factor (Bornovalova, Choate, Fatimah, Petersen, & Wiernik, 2020).

In multidimensional IRT models, Bi-factor confirmatory factor models have been influential in research on cognitive abilities because they often better fit the data than correlated factors and higher-order models (Morgan, Hodge, Wells, & Watkins, 2015).

**Purpose**

This study will focus on the assumptions of normality in the Bi-factor model with the graded response, preparing for future research. It is an extension of previous studies focused on unidimensional IRT models (DeMars, 2012) and multidimensional IRT models (Wang, 2018). Compared to other research designed for normality violation on one dimension or multi-dimension, this study uses bifactor IRT to check how the skewness of the general factor and specific factors will affect the recovery of parameters along with the severity of skewness and other dimensions, including sample size, the number of factors, items per factor. The good recovery of

**Model and Definition**

**Model Description**

MGRM, similar to GRM, describes the interaction between individuals and items for a wide range of psychological, educational, and medical outcomes measured with rating scales (Samejima, 1969; Bolt and Lall, 2003; Scherbaum et al., 2006).

Item 1

…

Item 10

Item 1

…

Item 10

Item 1

…

Item 10

Bifactor-IRT

In traditional GRM, all the possible responses to item j can be classified into a certain limited number of categories arranged in the order of attainment or intensity (Samejima, 1969). The probability that an examinee’s response falls at or above a particular ordered category given θ.

(1)

, (2)

Where D = 1.0 or 1.7 and k = 1, …, K. is the discrimination parameter for item j, denotes the item boundary parameter and θ denotes the latent trait of interest (Wang, 2018).

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,

- , (3)

In this study, we focus on the association between the constructs of psychometric tests, in which items are always linked to multiple latent traits. Hence, we extend the unidimensional GRM model to the multidimensional GRM (MGRM). in GRM is changed to , a vector of length H representing the latent traits of interest, the multidimensional version of probability equation is

, (4)

In which . This parameterization with D=1 is consistent with flexMIRT’s (Cai, 2015) default parameterization, serving as the “intercept”.

**Estimation of Non-Normality**

***Skewness in Simulation***

Fleishman (1978) noted that real-world distributions of variables are typically characterized by their first four moments (i.e., mean, variance, skew, and kurtosis). Tadikamalla (1980) proposed 5 methods for generating non-normal random numbers. He compared these five methods and Fleishman’s for speed, ease of implementation, and versatility. Fleishman's method was fastest and easiest to implement, but lacked a distribution function, making certain calculations like percentiles impossible. A method used in this study (selected in fungible package) is based on an extension of Fleishman's [1978] power function method for the univariate case.

In the univariate case, to defining a variable, Y, as a linear combination of the first three powers of a standard normal random variable X:

Y = a + bX + cX2 + dX3. (6)

The constants Mean, variance, skewness, and kurtosis (notated as a, b, c, and d) are chosen to provide Y with the specified distributional form.

b2 + 6bd + 2c2 + 15d2 -- 1 = 0 (7)

2c(b2 + 24bd + 105d2 + 2) – γ1 = 0 (8)

24[bd + C2(1 + b2 + 28bd) + d2(12 + 48bd + 141c2 + 225d2)] – γ2 = 0, (9)

where γ1 is the desired skewness and γ2 is the desired kurtosis, a is determined from

a = - c (10)

Although some researcher used skewness and kurtosis together to estimate the non-normality n in their research. We still cannot find many instructors say much about kurtosis, because it is difficult explain and judge from histograms. The more important thing is that data sets containing extreme values will not only be skewed, but also generally will be leptokurtic. some non-normal skewness is separable from non-normal kurtosis (Doane & Seward, 2011). The study will examine the impact of skewness while maintaining kurtosis at a constant value of 7, and manipulate three levels of skewness (normal=0, moderate=1.5, severe=3) for both the general and specific factors.

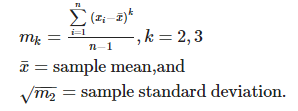
***Skewness in Estimation***

Non-normality of density appears both in raw scores and latent traits when researcher collect, and analysis reported questionnaire or scale from patients and other participants. In the graphic study and model estimation, skewness, and kurtosis of X (raw scores) and theta (latent traits) are conveniently used to measure and show. Skewness of a probability distribution refers to the departure of the distribution from symmetry. A symmetric distribution has no skewness, a distribution with longer tail on the left is negatively skewed, and a distribution with longer tail on the right is positively skewed (Sharma, Kumar & Chaudhary, 2009).

Mathematicians address skewness in terms of the second and third moments around the mean, and a few software packages (e.g., Stata, Visual Statistics, early versions of Minitab) report the traditional Fisher-Pearson coefficient of skewness, g1:

(11)

In which,

 (12)

(13)

Meanwhile, other software packages available to educators (e.g., Minitab, Excel, SPSS, SAS) include an adjustment for sample size, and provide the adjusted Fisher-Pearson standardized moment coefficient:

(14)

By comparing bias and mean squared error (MSE) of different measures of skewness in samples of various sizes from normal and skewed populations, G1 is shown to perform well (Joanes & Gill, 1998). Serval R package can offer a simple method to calculate the skewness. In this study, fungible package, applying G1 as method to calculate skewness, will be used to estimate skewness of predicted latent traits, noted as .

Instead of comparing mean of skewness of predicted theta with setting skewness (0, 0.4, 0.8), we compute the mean of skewness of generated theta (real theta) to calculate the bias between skewness of predicted theta and skewness of generated theta.

**Simulation**

**Design**

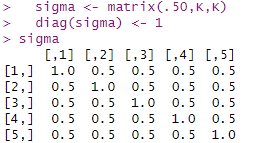
This study conducts a Monte Carlo simulation study of the bifactor model with one general factor and three specific factors, using the manipulated factors that have been implemented in previous research (Wang et al., 2018; Svetina et al., 2017; ), including sample size (two levels), test length (two levels), intercorrelation (two levels), and the number of non-normal dimensions (2 levels at general factor and 4 levels at specific/group factor; see Table 1). Degree of nonnormality is fixed at skewness 0.2 and kurtosis 7.0 (Wang et al., 2018).

Table 1 Simulation Design

|  |  |  |
| --- | --- | --- |
| Manipulated parameters | Number of levels | Values of levels |
| **Data Structure** |  |  |
| Sample size (N) | 3 | N = 250, 500, 1000 |
| Item per Factor (J) | 2 | J = 5, 10 |
| Factor (F) | 3 | F = 2, 3, 4 |
|  |  |  |
| **Skewness of Theta** |  |  |
| Skewness on general factor (GF) | 3 | GF = 0, 0.4, 0.8 |
| Skewness on specific/group factor (SF) | 3 | SF = 0, 0.4, 0.8 |

**Item parameter**

In psychological and psychiatric research, discrimination of general factor is commonly positive and not greater than 2.88 (Berkeljon,2012; Raines,2015; Atkinson,2018; Auné,2020), which is generally closed to [1.1, 2.8], which was applied by Rijmen’s (2011) simulation research. In our simulation phrase, discriminations of general factor are normally distributed in [1.1, 2.8], noted as ag. In previous research, discriminations of specific factor, aS are always smaller than the discrimination of general factor in same item. [0, 1.5] is a reasonable interval for simulation (Wang, 2018), and was applied in this study. Correlation between dimensions is all set as 0.5.



Item difficulty values can theoretically range from minus infinity to infinity, values typically vary from -2 to +2 (Hambleton, 1993; Hambleton & Swaminathan, 1985).In an application of the Argentine Version of the UCLA-LS (UCLA Loneliness Scale), there are three thresholds, which are b1[-1.40, 0.5], b2 [0.25,1.86], b3 [1.00, 3.63] (Auné, Abal, & Attorresi, 2020). Four-points Likert scale is prevalently employed in psychological and psychiatric test to analysis participants’ or patients’ latent traits or personalities. This study generated normally distributed thresholds, b1[−2, −0.67], b2[−0.67, 0.67], and b3[0.67, 2], for three thresholds to distinguish the possibilities of choosing each item.

**Person parameter**

In bifactor model, each object (participant patient) has one latent trait on general factor (θg) and several latent traits on specific factors (θs), in which k is the number of specific factors. There are three levels of skewness (normal=0, moderate=1.5, severe=3) were manipulated for each general factor and specific factors. This study maintain kurtosis at a constant value of 7. Although general latent trait and specific latent traits have different discrimination ag, no previous research focused on different algorithms’ performances on recovery of θg and θs. Here, different severity of skewness will be applied on two latent traits of single individual. This study set three level of skewness as 0, 1.5, 3, separately. There were nine combinations of non-skewed, moderately skewed, and severely skewed. All latent traits on specific factors (θs) are set equally. Correlation among all factor are set as 0.5.

**Estimation**

A fully crossed design for all these manipulated factors yielded a total of 162 conditions, each of which was replicated 30 times using packages and code written in R (R Core Team, 2021). Full-information maximum likelihood estimation in FlexMIRT is the only method utilized in this paper, the first step of the whole research. To increase the convergence speed in FlexMIRT, the number of integration quadrature points was reduced to 21 (default is 49), and the range was set to -3.5 to 3.5, which is also applied in Wang’s research (2018). Because FlexMIRT only offer cjk instead of bjk, we use bjk = -cjk/aj to transit cjk to bjk.

***Replications***

In a pilot study conducted in which the simulation procedure was implemented on 36 conditions for 50 replications, Wang (2018) found that the recovery indices stabilized after 30 replications. According to this result, we decided to replicate each of the 162 conditions in the study for 30 times. Means and standard deviations of each dependent variable were computed across replications.

***Evaluation criteria***

Performance was evaluated by computing the average bias and root means square error (RMSE) for the parameter estimates across 30 replications within each condition. Bias was defined as the average difference between the estimated and true values of the parameters across J items, while the RMSE was obtained by taking the square root of the mean of squared deviations of estimated parameter values about their true values.

The relative bias is estimated for all the parameters of model (ag, as, b1, b2, b3) as,

(9)

In which, is the predicted parameters () across valid replications and is the real parameters (agj, asj, b1j, b2j, b3j) simulated at the beginning of this research. In the , *j* is the order of items in each condition, from 1 to N, N calculated by number of items in each specific factor multi by number of specific factors. In equation, the number of replications for each condition is from 1 to M. In this study, M is fixed at 30.

The RMSE of

. (10)

The RMSE depends on the balance between the bias and standard deviation or variance. To estimate how skewness of density of real persons’ latent traits θg and θs straightforwardly, this study will estimate the variance of these parameters, instead of RMSE.

A five-way analysis of variance (ANOVA) for Bias, RMSE, and Variance were employed to compare and summarize the results of this simulation study.

**Results**

To check the performance of flexMIRT dealing small sample size, 50 was applied but had a poor convergence rate. Hence, we delete sample size 50 and all across conditions in this research.

**Skewness Estimation**

In simulation, we manipulated the skewness of real theta (θ) to generated multiple groups of theta (θ\*) with three conditions for both of general factor θg and θs. To avoid the error from simulation, we used the same method (R package DescTools) to calculate and compare G1 skewness for both simulated θ\*and predicted theta , following the equation 7.

**Table 1.** Observed skewness of distributions of simulated θ\* and predicted in General Factor and Specific Factor, in condition (Factor (F) = 4, Item per Factor (J) = 10, Sample Size = 1000)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Real Theta θ |  | Generated θ\* | |  | Predicted theta | |
| (General Factor-Specific Factor) |  | General Factor | Specific Factor |  | General Factor | Specific Factor |
| Normal-Normal (0,0) |  | -0.01 | 0.01 |  | -0.02 | 0.00 |
| Normal-Moderate (0,0.4) |  | 0.01 | 0.39 |  | 0.01 | -0.03 |
| Normal-Severe (0,0.8) |  | -0.01 | 0.81 |  | 0.00 | -0.04 |
| Moderate-Normal (0.4,0) |  | 0.39 | 0.01 |  | 0.17 | -0.01 |
| Moderate-Moderate (0.4,0.4) |  | 0.38 | 0.39 |  | 0.18 | -0.01 |
| Moderate-Severe (0.4,0.8) |  | 0.41 | 0.78 |  | 0.19 | -0.03 |
| Severe-Normal (0.8,0) |  | 0.81 | -0.02 |  | 0.42 | 0.00 |
| Severe-Moderate (0.8,0.4) |  | 0.79 | 0.40 |  | 0.41 | 0.01 |
| Severe-Severe (0.8,0.8) |  | 0.79 | 0.76 |  | 0.41 | -0.03 |

For the condition of Factor (F) = 4, Item per Factor (J) = 10, Sample Size = 1000, G1 method of calculating skewness offered a good performance on estimating the skewness of the simulated θ\* compared to the real theta, which also showed that the simulation with 30 replications is reliable to generate representative latent traits both in general and specific factors.

According to Table 1, there are different performance on the estimation of skewness of distributions for both predicted general-factor theta (θg\*) and specific-factor theta (θs\*).The skewness of general-factor theta dropped around half compared to the θg\*, while all the skewness of specific-factor theta were closed to 0. In conclusion, the skewness of distribution of general-factor theta can be partially identified in the model estimated by the FIML algorithm (flexMIRT), but all distribution in specific-factor theta is normal (non-skewed). In flexMIRT, specific-factor latent trait

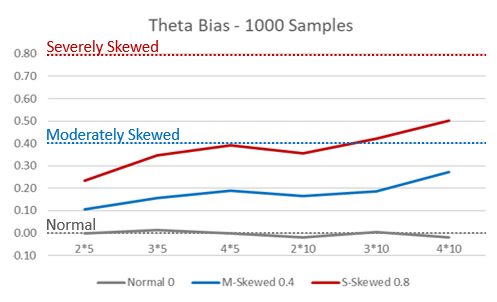


Figure 1. Skewness of predicted theta (1000 Sample Size)

In Figure 1, X-axis is factor multiple items per factor, and Y-axis is skewness scale. Solid lines are skewness of predicted theta of three skewed level (normal, moderate skewed, severe skewed) in different combinations of factor number and item number in each factor; the dashed lines are real theta in three skewed setting. We can address that there is a big underestimate in both moderate skewed and severe skewed situations. When we compare 4\*5 and 2\*10 conditions, both having 20 items totally, higher number of factor (4 Factor \* 5 Items) shows a better performance in skewness detection.

**Item parameter recovery**

When we used repeated-measures ANOVA, main effects and interactions between factors (number of factor, number of items per specific factor, sample size, skewness of theta on general factor, and skewness of theta on specific factor) explained very little variance of the all parameters estimates, including aG, aS, b1, b2, and b3(R-squaredag=0.01, R-squaredas=0.01, R-squaredb1=0.01, R-squaredb2=0.01, R-squaredb3=0.01,).

**Table 2.** Source of variance in 5-way ANOVA to estimate bias of parameters, effect size ω² (partial) greater than 0.001

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | F | p | ω² (partial) |
| ag | Sample Size | 408.553 | < 2e-16 | 7.41E-03 |
| Item per factor | 175.36 | < 2e-16 | 1.60E-03 |
|  |  |  |  |  |
| as | Number of Factor\* Sample Size | 45.898 | < 2e-16 | 1.64E-03 |
| Item per factor | 151.241 | < 2e-16 | 1.37E-03 |
| Sample Size | 75.003 | < 2e-16 | 1.35E-03 |
|  |  |  |  |  |
| b1 | Sample Size | 201.443 | < 2e-16 | 6.48E-03 |
| Item per factor | 356.977 | < 2e-16 | 1.83E-03 |
|  |  |  |  |  |
| b2 | Skewness of theta on general factor | 239.311 | < 2e-16 | 4.35E-03 |
|  |  |  |  |  |
| b3 | Sample Size | 164.396 | < 2e-16 | 6.68E-03 |
| Item per factor | 368.043 | < 2e-16 | 1.49E-03 |

In ag parameter, the most important source of variance was produced by sample size, F (2, 108973) =408.553, p < .001), ω² (partial)=0.007. In as parameter, the most important source of variance was produced by number of items per specific factor, F (1, 108973) = 151.24, p < .001), ω² (partial)=0.001. In b1 parameter, the most important source of variance was produced by sample size, F(2, 108973) =356.97, p < .001), ω² (partial)=0.006. In b2 parameter, the most important source of variance was produced by skewness level of theta on general factor, F(2, 108973) =239.31, p < .001), ω² (partial)=0.004. In b3 parameter, the most important source of variance was produced by sample size, F(2, 108973) =368.04, p < .001), ω² (partial)=0.006. In repeated measures ANOVA, this result is not offer a solid conclusion on the effect size of manipulated conditions.

In the results of flexMIRT, some predicted parameters have very extreme value, such as predicted aj = 117.54, far away from the real value 2.06, which is not a single phenomenon in 109136 item’s predicted parameters (both in aG, aS, b1, b2, b3).

**Discussion**

This simulation study addresses one primarily different result compared to previous research: Skewness on general factor and specific factor are separately manipulated and analysis their effects on model estimation. The other result is that we collected the conditions from practical psychological and psychometric research, offering a wide covered variety of conditions, which can offer a useful suggestion for subsequent research. The conditions are combined of factor number (2,3,4), items per factor (5,10), sample size (250,500,1000), skewness on G-factor (0,0.4,0.8) and skewness on S-factor (0,0.4,0.8). In total, there are 3\*2\*3\*3\*3 = 162 conditions are manipulated. Based on the previous research (Wang, 2018), the replication of each condition was set 30 in this research.

In this phrase of study, we compared the performance of flexMIRT in predicting the skewness of theta across various conditions We use same equation to calculate and compare skewness of and θ\* to avoid the bias of simulation, only focus on the bias of model estimation. In G-factor, predicted latent traits of objects are closed to half of the real/simulated latent traits of object (θg\*). In S-factor, predicted latent traits of objects are almost closed to 0, far away from the real/simulated latent traits of object (θS\*). In conclusion, flexMIRT can estimate the skewness on general factor, both in high-skewed (0.4) and moderate-skewed (0.8), but cannot offer a good estimation on specific factor.

A repeated-measures ANOVA are applied to compare all factors in simulation and their interactions. In some relatively small sample size condition, the flexMIRT cannot get convergence in each replication. We will raise the replication number in the next step. In the model estimation, flexMIRT output very extreme value on some predicted parameters, such as predicted aj = 117.54, compared to real value 2.06. Checking the simulation data of IRT and raw theta of object, we cannot offer a clear expression on this situation.

A set of functions are programmed in R environment for easily developing the next-step study.

In subsequent study, we will extend our estimate methods from flexmirt to Mplus and open-source packages with different estimation methods to compare the different performances on the estimation of parameters of Bi-factor IRT models.

In this study, the correlation of Factors in Bifactor are not considered. In Bifacotr-ESEM and Bifactor-BSEM model, the correlation can be manipulated in simulation and analysis the effect on model estimation.

**Reference**

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real 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.

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).

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.

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.

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 bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, *3*(1), 2-20.

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

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.

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 nonnormality. *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.

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.

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.

