**Examining the Impact of Nonnormality on Parameter Estimation of Bifactor Graded Response Model**

**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 nonnormality 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 nonnormality 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 nonnormality in the context of item response theory (IRT) models (Svetina et al., 2017; Woods, 2014), particularly using the bifactor IRT model. Bifactor model has been gaining popularity in psychological and other social sciences because of its flexibility to incorporate a general factor and some specific factors for the multidimentional latent factors. To the best of our knowledge, no previous study examines the impact of nonnormality on Bi-factor 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-GRM to check 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.

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. For example, the marginal maximum likelihood method is the most widely used approach for item parameters and to integrate out the person parameters, one often assumes a normal distribution or multivariate normal distribution. For person parameter estimation, maximum a posteriori (MAP) estimation has been shown to achieve precise estimation with fewer items than the ML method, 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, including item discrimination, threshold, and personal abilities on both general factor and specific factor.

The current study investigates the violation of normality assumptions in multi-dimensional models using Bifactor Graded Response Model (Bifactor-GRM). It examines 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 research explores various design factors, including the severity of skewness in the general and specific factors, sample size, number of factors, and number of items per factor.

**Theoretical Framework**

**Grade Response Model**

The Graded Response Model (GRM), a component of the broader Item Response Theory (IRT) used in psychometrics, is a model specifically tailored for ordinal responses, such as Likert scales. GRM is effective for predicting the likelihood of a respondent selecting a specific response level or higher on a survey (Samejima, 1969; Baker & Kim, 2004).

**Bifactor Grade Response Model**

The Bifactor Graded Response Model (Bifactor-GRM) is an extension of the conventional Graded Response Model (GRM) and is a part of Item Response Theory (IRT). 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 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.

where P is the probability to provide a response equal to k or greater given a person's location on G and a specific trait S, category k's item-intercept as defined , and the conditional item slope parameter on G () and on S (). The person parameter represents person i's location on G, whereas represents person I’s location on S. For each person we have a number of specific trait scores equivalent to the number of specific traits defining the model (Toland et al., 2017).

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

**Method**

**Design Factors**

This study is a Monte Carlo simulation study of the bifactor model with one general factor and two, three or four specific factors (Fs = 2, 3, 4), using the manipulated factors that have been implemented in previous research (Auné,2020; Mao,2022; Rijmen,2011; Svetina et al., 2017; Wang et al., 2018), including sample size (three levels: N= 250, 500, 1000 ), number of item per factor (two levels: I = 5, 10), and the degree of nonnormality on population’s latent traits (three levels at general factor and three levels at specific/group factor; see Table 1).

In bifactor models, each subject has one general factor (θg) and several specific factors (θsk), in which k is the number of specific factors. We simulate three levels of nonnormality on general factor and specific factors: normality (with skewness and kurtosis values of 0), moderate nonnormality (skewness: 2, kurtosis: 7), and severe nonnormality (skewness: 3, kurtosis: 21). There are nine different combinations of nonnormality for the general factor and specific factors. All the nonnormalities of the latent traits pertaining to the specific factors (θsk) are set to the same values in both 2, 3, and 4 specific factors settings.

**Table 1 Simulation Design**

|  |  |  |
| --- | --- | --- |
| Design factors | Number of 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 |

3 × 2 × 3 × 3 × 3 = Each of these conditions was replicated 500 times using the R package "SimMultiCorrData" in R (R Core Team, 2021).

**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), used as specific factor in this study. 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 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**

Blanca (2013) showed the measures of cognitive ability and of other psychological variables were included. The results showed that skewness ranged between −2.49 and 2.33. The values of kurtosis ranged between −1.92 and 7.41. The values for asymmetry and kurtosis between -2 and +2 are considered acceptable in order to prove normal univariate distribution (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. Since, we simulate three level of nonnormality, normality (skewness: 0, kurtosis: 0), moderate nonnormality (skewness: 2, kurtosis: 7), and severe nonnormality (skewness: 3, kurtosis: 21).There were nine combinations of non-skewed, moderately skewed, and severely skewed. 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 2000 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. These measures are calculated for both the two discrimination parameters, the three boundary parameters, and two personal parameters, for each replication.

**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 is the estimated parameters () across valid replications and is the true parameters (agj, asj, c1j, c2j, c3j , θgj, θsj,).

**Correlation**. Correlation measures the strength and direction of a linear relationship between two variables. It is a statistical measure that ranges from -1 to 1. In this study, a positive correlation (closer to +1) indicates that when true personal traits (θ) increases, the other tends to increase as well. A correlation close to 0 indicates a weak or no linear relationship between the variables.

To aid in comparing the effects and interactions among the manipulated variables, we conducted ANOVA analysis and effect size (η2) was computed.

**Preliminary Results**

**Item Parameter Estimation**

When analyzing item parameters, none of the interaction terms had an effect size larger than 0.05. We focused on ag (discrimination on the general factor), as (discrimination on the specific factor), and three locations c1, c2, and c3. In the bias test, we discovered that as the skewness and kurtosis of the population's general factor increased, the bias in estimating ag grew significantly. However, the bias in as estimation was not impacted. When the skewness and kurtosis of the population's specific factor increased, there was a slight increase in the bias of estimating as, but it did not affect the estimation of ag. For estimating the location parameter c, we took an average of c1, c2, and c3 instead of treating them separately. The results showed that the skewness and kurtosis of the population's general factor positively influenced the estimation of c, while the non-normality of the population's specific factor, sample size, and item number per factor had limited impact.

Regarding RMSE estimation, as the skewness and kurtosis of the population's general factor increased, the RMSE of estimating ag became noticeably higher. However, the increase in skewness and kurtosis of the population's specific factor had an imperceptible effect on the RMSE of estimating as. Item number per factor and sample size effectively impacted as. Sample size emerged as a major factor influencing all item parameters, including ag, as, and c.

**Personal Parameter Estimation**

The choice of algorithm used for estimating theta plays a significant role in personal parameter measurement. This is especially true for theta related to general factors, as it can introduce bias, increase the root mean square error (RMSE), and impact the correlation between estimated theta and the true theta. However, the algorithm mainly affects the RMSE of theta related to specific factors, with no significant impact on bias and only a slight influence on correlation.

When there is a greater deviation from normality in both general and specific factors, the bias and RMSE in estimating theta for these factors separately become more pronounced. Additionally, the correlation between the population's theta for general factors and the true theta decreases.

The number of specific factors can affect the bias, RMSE, and correlation in estimating theta for general factors, as well as the bias and RMSE in estimating theta for specific factors.

Another factor that increases the RMSE in estimating theta for both general and specific factors is the sample size. At the same time, it decreases the correlation between the estimated theta and the true theta for specific factors.

**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 nonnormality 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 nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 87-107.

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 nonnormality. *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 nonnormality 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 nonnormality on some multivariate tests and robustness to nonnormality 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 nonnormality. *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 nonnormality? 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 nonnormality 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 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.

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 nonnormality 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.