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

**Purpose**

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. This study will focus on the assumptions of normality in the Bifactor 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.

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, involving 162 conditions (500 replications per condition). The precision of estimation can be influenced by the skewness and kurtosis of population’s distribution of latent traits. Specifically, the skewness and kurtosis of latent traits within the specific factors (θs) have a lesser impact on the estimation of item parameters and personal parameters compared to the general factor (θg). To compare the bias, RMSE (Root Mean Square Error), and correlation between estimated parameters and actual parameters across different conditions, a repeated-measures ANOVA was performed. The results indicated that three major factors impacting the estimation of item parameters and individuals' latent traits (theta) are sample size, the number of items per factor, and the skewness and kurtosis of the general factor.

**Theoretical Framework**

**Model Description**

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.

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,

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

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:

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

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

Our research focuses on examining how the distribution of latent traits is influenced by skewness and kurtosis when estimating parameters of a bifactor model. We aim to investigate the impact of nonnormality on the estimation process. To achieve this, we simulate three levels of nonnormality: 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 (θs) are set to the same values in both 2, 3, and 4 specific factors settings.

***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:

In which,

, and

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:

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 .

**Method**

**Design**

This study conducts a Monte Carlo simulation study of the bifactor model with one general factor and two, three or four specific factors, using the manipulated factors that have been implemented in previous research (Rijmen,2011; Svetina et al., 2017; Wang et al., 2018; Auné,2020; Mao,2022), including sample size (three levels: N= 250, 500, 1000 ), 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).

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 (I) | 2 | I = 5, 10 |
| Factor (F) | 3 | F = 2, 3, 4 |
|  |  |  |
| **Nonnormality of Latent Traits (Theta)** |  |  |
| Nonnormality on general factor (Fg) | 3 | Normal: Skewness = 0, Kurtosis = 0  Moderate: Skewness = 2, Kurtosis = 7  Severe: Skewness = 3, Kurtosis = 21 |
| Nonnormality on general factor (Fs) | 3 | Normal: Skewness = 0, Kurtosis = 0  Moderate: Skewness = 2, Kurtosis = 7  Severe: Skewness = 3, Kurtosis = 21 |

**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 phase, the discriminations of the general factor (referred to as "ag") are normally distributed within the interval of 1.1 to 2.8. Previous research has consistently shown that the discriminations of the specific factor (referred to as "aS") are always smaller than the discrimination of the general factor within the range of 0 to 1.5, as supported by Wang (2018). In the bifactor model, the relationship between the general factor and specific factor is considered orthogonal, meaning that the correlation between them is zero. Consequently, the latent traits associated with the general factor and specific factor are simulated separately.

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 (locations) to distinguish the possibilities of choosing each item.

**Person parameter**

In bifactor model, each object (participant or patient) has one latent trait on general factor (θg) and several latent traits on specific factors (θsk), in which k is the number of specific factors. There are three levels of nonnormality (normal, moderate, and severe) were manipulated for each general factor and specific factors. During our research process, we discovered that the interaction between skewness and kurtosis has a significant influence on the degree of nonnormality, rather than considering skewness or kurtosis individually. This combined effect of skewness and kurtosis plays a dominant role in determining the Kullback-Leibler divergence, also known as relative entropy or I-divergence. The Kullback-Leibler divergence, denoted as DKL(P‖Q), serves as a statistical measure to quantify the dissimilarity between two probability distributions: P, which is normally distributed, and Q, representing a nonnormal distribution characterized by specific skewness and kurtosis values.

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**

A comprehensive experimental design was employed, incorporating all the manipulated factors, resulting in a total of 162 unique conditions. Each of these conditions was replicated 500 times using the R package "SimMultiCorrData" in R (R Core Team, 2021). The item parameters in this study were estimated using the "bfactor()" function from the R package "mirt", with 2000 iterations. For estimating the personal parameters, two methods, namely MAP (Maximum A Posteriori) and ML (Maximum Likelihood), were utilized. Within the R package "mirt," the estimation of personal parameters involved utilizing the "fscores()" function. In this package, the thresholds or locations are calculated as cjk, as described in equation (6).

***Replications***

In order to ensure reliable recovery indices, we opted to replicate each of the 162 conditions in the study 500 times. This approach allowed us to calculate the means and standard deviations of each dependent variable by aggregating the data across these replications.

***Evaluation criteria***

Performance was evaluated by computing the average bias and root means square error (RMSE) for the parameter estimates across 500 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, c1, c2, c3) as,

In which, is the predicted parameters () across valid replications and is the actual parameters (agj, asj, c1j, c2j, c3j) 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 500.

The RMSE of

The RMSE depends on the balance between the bias and standard deviation or variance. To estimate how skewness of density of actual 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 Correlation were employed to compare and summarize the results of this simulation study.

**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 actual 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 actual 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 actual theta for specific factors.

**Reference**

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

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.

**Appendix**

Table 1. Effect sizes (>.05) of main effects and interactions for the bias and RMSE of estimation on item parameters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Statistic and Source of Variation | a.G | a.S | c1 | c2 | c3 |
| **Bias** |  |  |  |  |  |
| Skewness on General Factor | **.857** | .001 | **.343** | **.621** | **.440** |
| Skewness on Specific Factor | .000 | **.214** | .022 | **.052** | .026 |
| Item Number per Factor | .001 | **.054** | **.090** | .018 | .022 |
| Sample Size | .020 | .021 | **.080** | .004 | **.051** |
|  |  |  |  |  |  |
| **RMSE** |  |  |  |  |  |
| Skewness on General Factor | **.409** | .001 | .039 | **.255** | .007 |
| Item Number per Factor | **.080** | **.236** | **.112** | .049 | **.103** |
| Sample Size | **.084** | **.234** | **.256** | **.264** | **.261** |

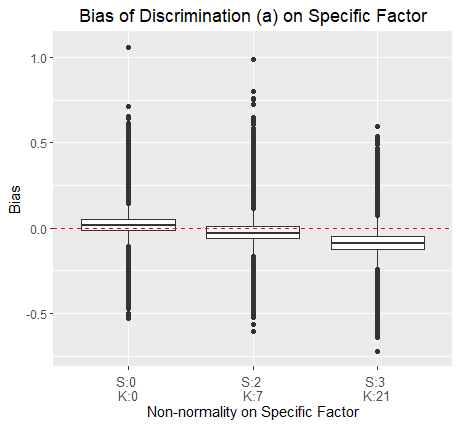
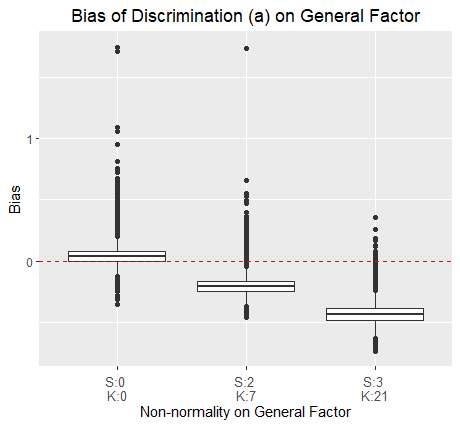


Figure 1. Bias of parameter "a" on the general factor (aG) and the specific factor (aS)

A picture containing text, diagram, screenshot, line

Description automatically generated

Figure 2. Bias of parameter c1, c2 and c3.

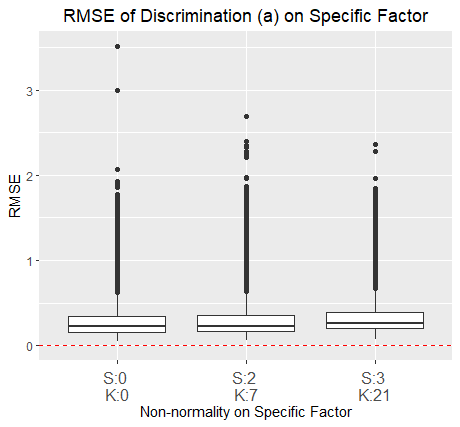
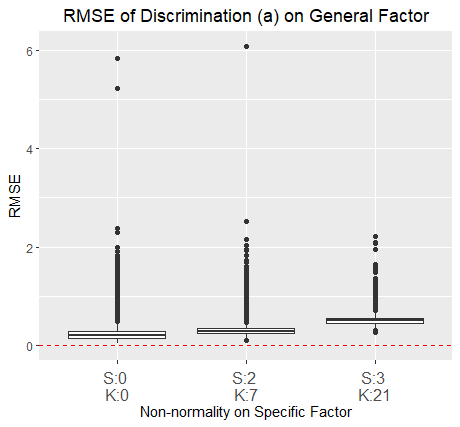


Figure 3. RMSE of parameter "a" on the general factor (aG) and the specific factor (aS).

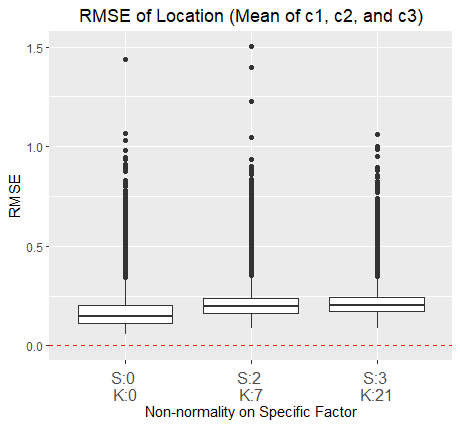


Figure 4. RMSE of parameter c1, c2 and c3.

Table 2. Effect sizes (>.05) of main effects and interactions for the bias of bias, RMSE and Correlation of estimation on personal parameters.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Bias** | | **RMSE** | | **Correlation** | |
|  | ThetaG | ThetaS | ThetaG | ThetaS | ThetaG | ThetaS |
| Factor | **.080** | **.155** | **.611** | **.700** | **.714** | .034 |
| I | .003 | .042 | **.596** | **.759** | **.713** | **.706** |
| N | .027 | .034 | **.068** | **.064** | .014 | **.134** |
| Fg | **.499** | .014 | **.269** | **.074** | **.481** | .020 |
| Fs | .001 | .009 | .003 | **.063** | .006 | **.221** |
| MAP | **.184** | .008 | **.906** | **.965** | **.771** | **.074** |
|  |  |  |  |  |  |  |
| Factor:I | .049 | **.060** | **.227** | **.087** | **.298** | .041 |
| Factor:N | .047 | **.050** | .090 | **.071** | **.157** | **.147** |
| I:N | .003 | .024 | .001 | .018 | .001 | **.080** |
| Factor:Fg | .004 | .007 | .001 | .000 | **.055** | .000 |
| I:Fg | .000 | .023 | .002 | .001 | **.092** | .001 |
| Factor:MAP | .037 | **.152** | **.284** | **.273** | **.095** | .001 |
| I:MAP | .009 | .038 | **.374** | **.662** | **.258** | .003 |
| N:MAP | .012 | .036 | **.059** | **.100** | .010 | .003 |
| Fg:MAP | **.066** | .004 | **.083** | **.059** | .013 | .001 |
| Factor:I:N | .031 | .023 | **.12** | .031 | **.126** | **.084** |
| Factor:I:MAP | .039 | **.065** | **.128** | .045 | .037 | .000 |
| Factor:N:MAP | **.057** | **.056** | **.064** | .032 | **.057** | .002 |
| Factor:I:N:MAP | .034 | .029 | **.059** | .017 | .040 | .001 |

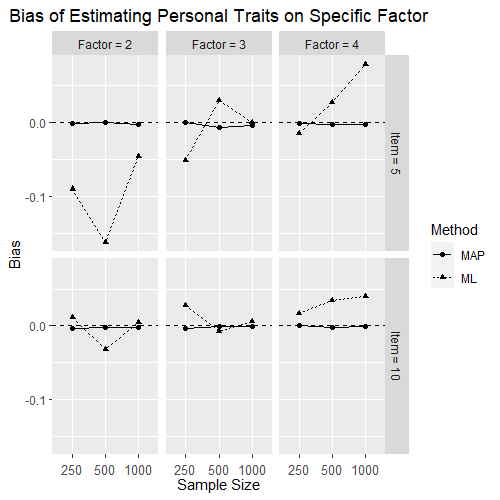
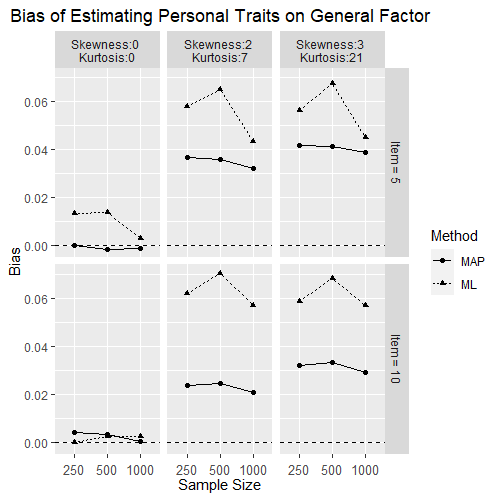
****

Figure 5. Bias of theta on General Factor and Specific Factor estimated by MAP and ML method.

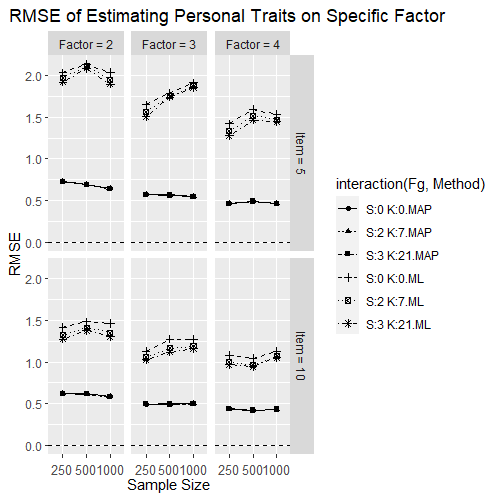
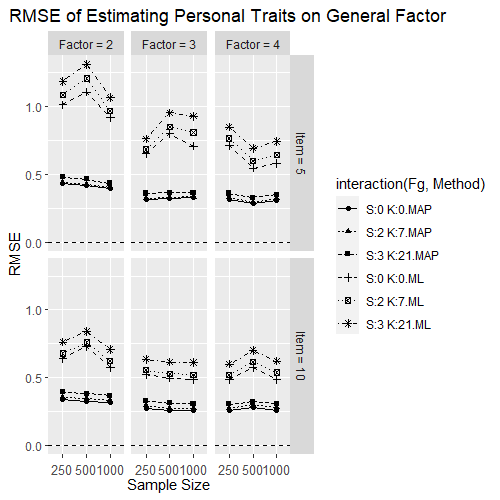
****

Figure 6. RMSE of theta on General Factor and Specific Factor estimated by MAP and ML method.

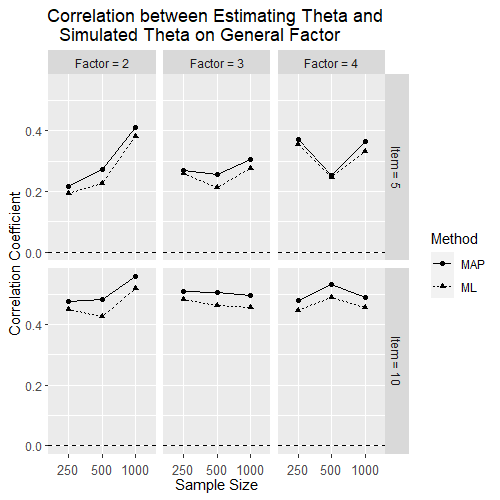
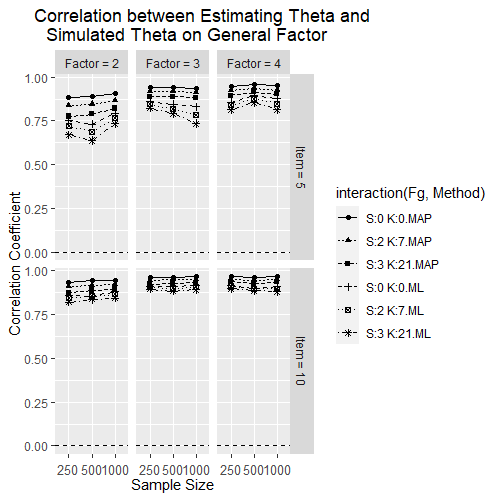
****

Figure 7. Correlation of theta on General Factor and Specific Factor estimated by MAP and ML method.