CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 7(3) (2022), Pages 332-344
p-ISSN: 2086-0382; e-ISSN: 2477-3344

A Monte Carlo Simulation Study to Assess Estimation Methods in Confimatory Factor Analysis (CFA) on Ordinal Data

Nina Fitriyati*, **and Madona Yunita Wijaya**

Universitas Islam Negeri Syarif Hidayatullah Jakarta, Indonesia

Email: nina.fitriyati@uinjkt.ac.id

ABSTRACT

Likert-type scale data are ordinal data and are commonly used to measure latent constructs in the educational, social, and behavioral sciences. The ordinal observed variables are often treated as continuous variables in factor analysis, which may cause misleading statistical inferences. Two robust estimators, i.e., unweighted least squares (ULS) and diagonally weighted least squares (DWLS) have been developed to deal with ordinal data in confirmatory factor analysis (CFA). In this research, we conduct an extensive simulation study to examine the impact of ULS and DWLS estimations in the accuracy of parameter estimates as well as the model fit measures by varying the number of Likert scale points. We use synthetic data generated in a Monte Carlo experiment to explore the behavior of these methods (DWLS and ULS) and compare their performance with normal theory-based ML and GLS (generalized least squares) under different levels of experimental conditions. The simulation results indicate that both DWLS and ULS yield consistently accurate parameter estimates across all conditions considered. The Likert data can be treated as a continuous variable under ML or GLS when using at least five Likert scale points to produce trivial bias. However, these methods generally fail to provide a satisfactory fit. Furthermore, we provide empirical studies in the field of psychological measurement data to present how theoretical and statistical instances have to be taken into consideration when ordinal data are used in the CFA model.

Keywords: confirmatory factor analysis; diagonally weighted least squares; generalized least squares; Likert data; maximum likelihood

INTRODUCTION

Factor Analysis such as CFA is often used in various fields, including social science, economics, and psychology to validate survey instruments. CFA is developed from the concept of Exploratory Factor Analysis (EFA) where researchers have the hypothesized model based on theory or previous analytical research. The model should be specified before analysis regarding the number of factors in the model, the number of indicators reflecting each factor, and whether a relationship exists between these factors. Factor analysis plays an important role to provide evidence of construct validity and information about the internal structure of the measurement instrument [1] [2]. Thompson [3] stated that CFA is also known as a measurement model which is an important component of the structural equation model (SEM) class, describing how the measured indicators can reflect a latent variable/construct. The measurement model describes how the latent variable depends on the indication of the observed variable which also explains the measurement properties such as the validity and reliability of the observed variable [4].

Submitted: December 20, 2021 Reviewed: March 07, 2022 Accepted: March 28, 2022
DOI: http://dx.doi.org/10.18860/ca.v7i3.14434

A questionnaire is one of the tools in quantitative research that is generally used to study latent constructs through their observed indicators. The indicators are collected through a series of questions in the form of a questionnaire measured on a Likert-type scale. This scale offers a series of fixed-choice options where there is a rank order between the choices given [5]. The results of measuring data using a Likert scale are also known as ordinal data.

When collecting data with a Likert scale for analysis, researchers tend to ignore the actual data structure and treat it like a continuous variable since the methods designed for continuous variables tend to be easy to apply and easy to use. In this case, the Maximum Likelihood (ML) method is used to estimate the model parameters [6]. The ML method is the most popular since it gives asymptotic unbiased results and consistent parameter estimates [7] [8] [9] [10]. However, this method uses Pearson correlation which requires the assumption of continuous measurement for both latent and observed variables. Pearson correlation also tends to underestimate the true correlation between ordinal variables. However, in principle, this method cannot be directly applied to ordinal data because the characteristics of the data itself tend to be different. Several theories and estimation methods are available which are designed to handle ordinal data, such as DWLS (Diagonally Weighted Least Squares) and ULS (Unweighted Least Squares).

Several studies have addressed the comparison between ULS and DWLS performance in finite samples [11] [12] [13] [14]. All these studies have certain limitations in that they did not compare with standard estimation procedures such as ML or GLS. As of today, many researchers are still in their comfort zone using the standard estimation, knowing that the observed variables violated the normal assumption. In Maydeu-Olivers [11] study only considered a non-standard model, whereas Tate [12] considered one replication per condition. The other two studies only considered a certain condition in terms of the number of categories used per indicator.

To conduct further research in this area, an extensive simulation study was conducted to examine the impact of ULS and DWLS estimations in the accuracy of parameter estimates as well as the model fit measures by varying the number of Likert scale points. In addition, ML and GLS were also compared to give the readers insight into the risk of treating ordinal data as a continuous variable in the factor analysis.

METHODS

This research was implemented as a Monte Carlo simulation study, where it involves creating data by pseudo-random sampling to study the behavior of statistical methods under various conditions [15]. In this study, we evaluate the performance of CFA estimation methods under different experimental conditions. In addition, numerical examples using secondary data are also considered to support the findings.

Estimation Methods

The most common method used for parameter estimation in the CFA model is the maximum likelihood (ML). It is used more frequently in many research as it has many desirable statistical properties, including asymptotic unbiasedness, asymptotic consistency, asymptotic normality, and asymptotic efficiency. The ML estimation method assumes that the observed indicators are multivariate normally distributed and measured on continuous scales [16]. To estimate the model parameters (θ), ML maximizes the likelihood function of the observed data, or equivalently minimizes the following function [7]:

$$F_{ML} = \log|\textstyle\sum(\theta)| - \log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\textstyle\sum(\theta)^{-1}) - k, \tag{1}$$

where S is the sample variance-covariance matrix, $\sum(\theta)$ is the model implied covariance matrix, and k is the number of observed indicators.

Generalized Least Squares (GLS) is another estimation method that assumes the multivariate normality of the data similar to ML. The model parameters are estimated by minimizing the following function:

$$F_{GLS} = \frac{1}{2} tr(\mathbf{S} - \textstyle\sum(\theta)\mathbf{S}^{-1})^2. \tag{2}$$

The main difference between the two estimators is that the sample covariance matrix is used in the weight matrix instead of the model-implied covariance matrix. However, this estimator provides a better empirical fit than ML when the models are misspecified [17] [18] [19].

The assumptions of normality and continuous scales are often violated since many research, such as psychology, social sciences, and educational research, use Likert scales to measure observable variables. Likert-type scales are categorical (ordinal) data where the response categories have a rank order, i.e., one score can be said to be higher than another but not the distance between the two scores. If the assumption is not satisfied, then the resulting CFA model may not be reliable and the conclusion obtained can be misleading.

Alternative estimators available are DWLS and ULS, which have been suggested to deal with ordinal data [20] [6]. When the data are treated as categorical (ordinal), correlation structure is involved instead of covariance structure and the model is fitted to polychoric correlations. Both methods share a similar fit function, i.e., least-square function to estimate the model parameters involving minimizing the following objective function:

$$F = \big(\mathbf{S} - \boldsymbol{\sigma}(\theta)\big)'\mathbf{W}\big(\mathbf{S} - \boldsymbol{\sigma}(\theta)\big), \tag{3}$$

where $\mathbf{W}$ is the weight matrix and $(\mathbf{S} - \boldsymbol{\sigma}(\theta))$ denotes the difference between the population threshold and polychoric correlations and those implied by the model. The weight matrix $\mathbf{W}$ determines the minimization procedure. When the choice of $\mathbf{W}$ is the identity matrix, the method is known as ULS (Muthen, 1993). A second choice is $\mathbf{W} = \big(diag(\boldsymbol{\Gamma})\big)^{-1/2}$ known as DWLS. It uses the elements in the diagonal matrix of the estimated polychoric correlations (the estimated variances) as the weights [21].

Simulation Design

Monte Carlo simulation studies were carried out to compare the impact of various estimation methods (ML, GLS, ULS, and DWLS) on the CFA model by manipulating four experimental factors, i.e., the number of latent variables or factors, number of items, the number of response categories, and sample sizes. More specifically, we are interested in evaluating four popular model fit indices (i.e., Chi-square test, RMSEA, CFI, and TLI) and also bias in factor loadings across different experimental conditions.

Different CFA models were generated under different settings of dimensionality, i.e., a one-factor model and a correlated two-factor model. Each factor was measured by four and eight ordinal observed indicators. At least four indicators are necessary for a one-factor model to be (over-)identified. Additionally, it was found to have optimal accuracy of parameter estimates and marginally improved by adding more indicators [22]. The coefficients of factor loadings on the same factor were varied from low to high loadings, i.e., 0.5 to 0.9, to mimic real-world data as it is very rarely happened to have constant factor loadings on the same factor. It was also suggested from many empirical research and simulation studies were reported standardized factor loadings range from 0.4 to 0.9 [23] [24] [25]. For a two-factor model, the inter-factor correlation was set to have a moderate correlation of 0.4. The factor variances were all set equal to 1 in the population. Four different sample sizes are considered from low to large samples: N = 100, 200, 500, and 1000 [26]. All generated models assumed symmetric observed distributions generated from normal latent response distribution. To investigate the influence of categorization or the number of Likert scale points (c), each indicator was generated by considering $c = 2,3,\dots,10$ categories.

A total of 144 experimental conditions $(2 \times 2 \times 9 \times 4)$ was created in this study by considering the combination of the four experimental factors, i.e., number of factors $(f = 1,2)$, number of items $(i = 4,8)$ number of categories $(c = 2,\dots,10)$, and sample sizes $(N = 100, 200, 500, 1000)$ [26]. Each experimental condition was evaluated by generating 500 datasets. All data were generated using Mplus while CFA model evaluations were carried out using the '*lavaan*' package in R [27] [28].

Outcome Variables

In this study, a model assessment was done on parameter estimates and model fit. For each experimental condition, parameter estimates including factor loadings and inter-factor correlation were examined. The bias, i.e., the difference between the estimated $(\hat{\theta})$ and true parameter (θ), were computed to evaluate the performance of the four estimation methods. The average absolute relative bias (ARB) was considered to take into account the magnitude of the true parameter value. Let $\hat{\theta}_{ij}$ be the parameter estimates of the jth parameter $(1,2,\dots,p)$ in the ith replicate $(i = 1,2,.\,.,R)$. The ARB can be expressed as follows [29]:

$$ARB = \frac{1}{R}\sum_{i=1}^{R}\left(\frac{1}{p}\sum_{j=1}^{p}\frac{(\hat{\theta}_{ij}-\theta_j)}{\theta_j}\right). \quad (4)$$

A trivial bias is indicated with an ARB value less than 5%, a moderate bias is indicated with an ARB value between 5% and 10%, and a substantial bias is indicated with an ARB value above 10% [30] [31].

To assess model fit, the Chi-square test and RMSEA were evaluated in terms of the rate of rejection (Type I Error) of the proposed model. A rate of rejection greater than 5% indicates inflated Type I error rates, showing that the test statistics may have been underestimated. Other fit indices such as TLI and CFI were evaluated by averaging across R replications. They were also computed as the proportion of satisfactory fit across R replications (TLI / CLI > 0.9).

RESULTS AND DISCUSSION

Parameter Estimation

The results of average relative bias (ARB%) values for factor loadings for each estimation method depending on the number of Likert scale points are presented in Figure 1 and 2. When ordinal indicators were treated as continuous, the number of Likert scale points had an impact on the accuracy of the estimated factor loadings, whether ML or GLS estimations were used. The estimated factor loadings suffered from a downward bias with a fewer number of Likert scale points ($c < 5$) and the bias increased dramatically when only considering two categories in the observed indicator variable. Increasing sample size did not seem to reduce the bias. Using at least five response categories produced moderate bias since the ARB values were less than 10%. The accuracy of parameter estimates improved with the increased number of response categories. The factor loadings were essentially unbiased with nine or ten response categories under ML and GLS.

The DWLS and ULS estimation methods were superior to the other two methods. Both methods consistently provided more accurate factor loadings, evidenced by their smaller ARB values. The resulting bias was trivial even in the condition of a small sample size ($N = 100$). Unlike ML and GLS, the number of Likert scale points did not influence the accuracy. The bias remained stable below 5% across a different number of response categories. In general, both methods appeared to perform well under various conditions.

The impact of the number of indicators to measure a latent construct on the accuracy of factor loading estimates can be notably seen in the one-factor model. The bias decreased as more indicators were used based on DWLS and ULS methods. In contrast, the ARB values were generally larger when using more indicators under ML and GLS. In a two-factor model, the bias was relatively similar whether 4- or 8- indicators per factor were used.

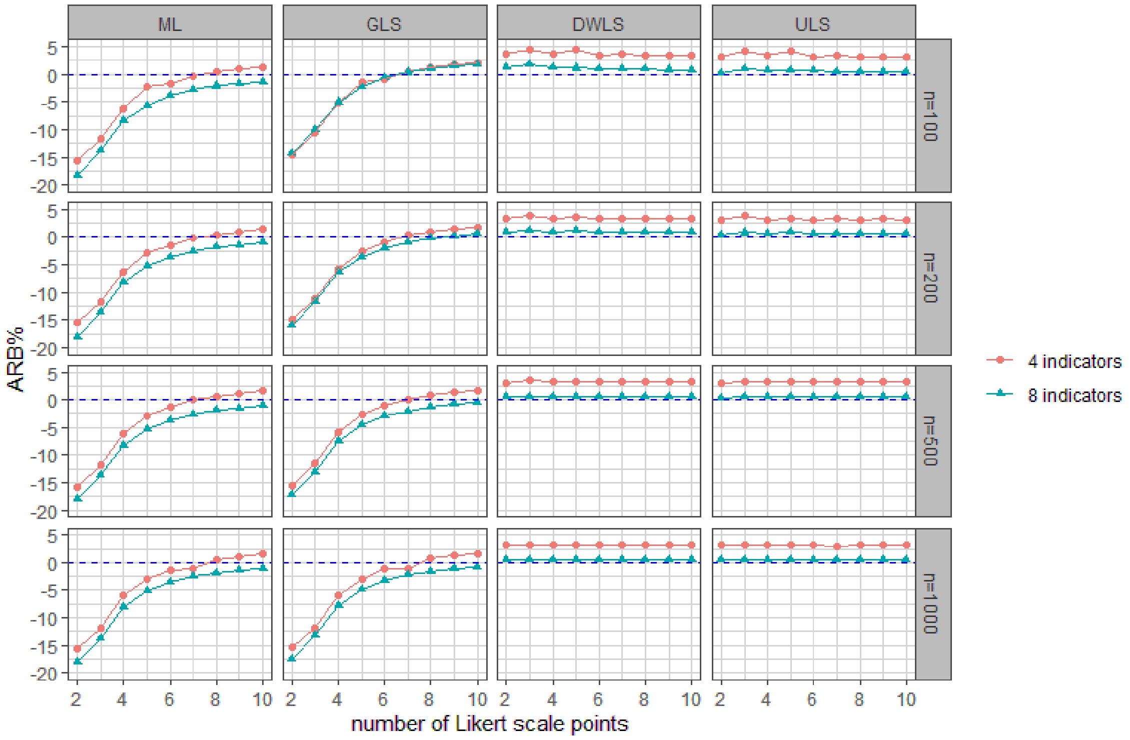

Figure 1. Plot of ARB(%) for factor loadings by varying the number of Likert scale points, the number of indicators per factor, and sample size, for a one-factor model.

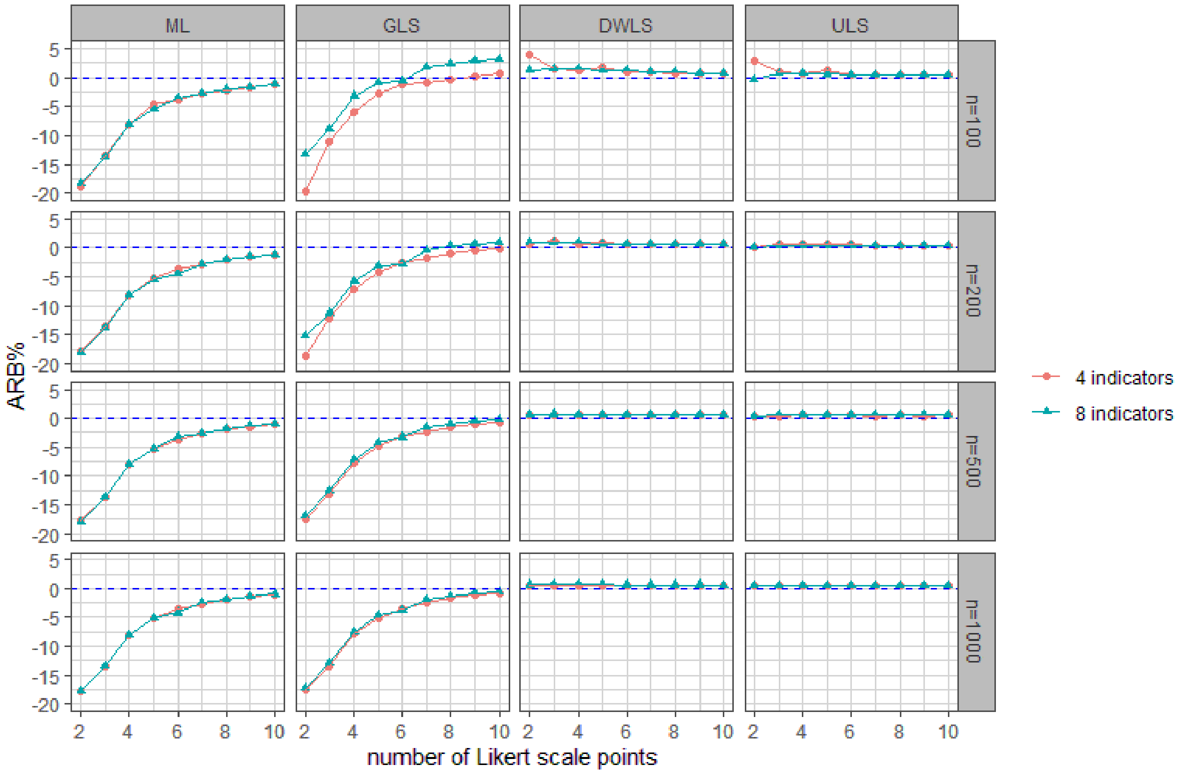

Figure 2. Plot of ARB(%) for factor loadings by varying the number of Likert scale points, the number of indicators per factor, and sample size, for a two-factor model.

The ARB values for inter-factor correlations in two-factor models are presented in Figure 3 by varying the number of indicators, the number of Likert scale points, and sample size. The plots also show a comparison of the performance between the four different estimation methods. Similar to factor loadings, the number of Likert scale points appeared to have an impact on the estimated inter-factor correlations. ML and GLS methods performed equally worst in estimating the correlation between the two factors in the condition when two-response categories were used since they substantially underestimated the true correlation parameter ($\text{ARB} < -10\%$). The ARB values can be improved by increasing sample size; however, the resulting estimates were still biased. In the case of five or more response categories, both methods were able to achieve accuracy in the inter-factor correlation estimates with trivial bias. Higher accuracy was obtained as the sample size increased. It is interesting to see that the ML estimator generally produced negative bias while in contrast, GLS produced positive bias. Meanwhile, the inter-factor correlation bias was consistently lower under DWLS and ULS and essentially unbiased across different scenarios. A larger sample size reduced relative bias, particularly in a two-factor model with four indicators per factor.

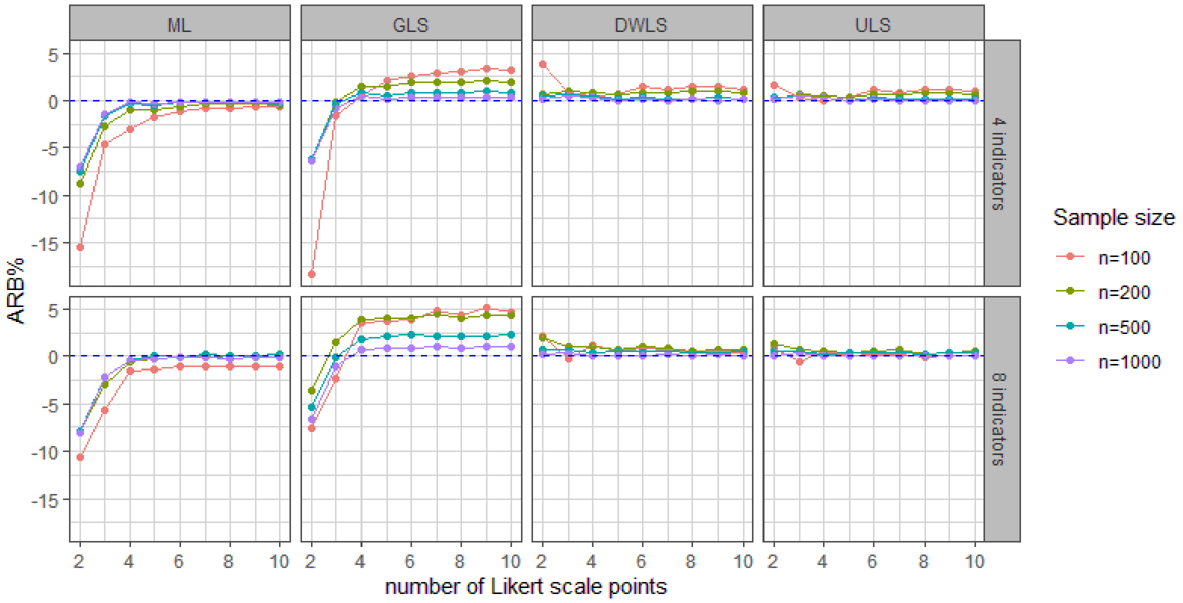

Figure 3. Plot of ARB(%) for inter-factor correlations by varying the number of Likert scale points and sample sizes for a two-factor model with four and eight indicators per factor.

Model Fit

Figure 4 and 5 present rejection rates associated with RMSEA and Chi-square test (note that not all experimental condition's results are shown here for brevity). The empirical type I error for the Chi-square test under ML and GLS estimation methods were well within the range of 0.03 and 0.10, which is very near to the nominal value of type I error ($\alpha = 0.05$) when considering four indicators per factor. The rate was the highest for a larger sample size, particularly for two response categories. Increasing the number of Likert scale points, the type I error was able to be controlled at a 5% level. In contrast, the type I error inflated along with the increase in the use of several indicators per factor, particularly for small sample sizes and under the ML method. GLS performed generally better than ML within this similar condition. Overall, DWLS worked well in maintaining Chi-square type I error rates across most experimental conditions. In contrast, the RMSEA fit indices were found to be insensitive to most conditions of the study. The rejection rates were consistently below 0.05, regardless of the estimation methods used. Except for the ULS method, which was found with poor performance in the condition of small sample size ($N = 100$) and two response categories.

The proportions of TLI and CLI values greater than 0.9 across 500 data generations obtained from fitting a one-factor model is displayed in Figure 6. Other models are not shown here since the resulting TLI and CLI was pretty much similar. The estimation methods did not influence both CFI and TLI values in larger sample sizes ($N \geq 500$), since they performed consistently well with a proportion above 0.9. In smaller sample sizes, GLS was the most conservative one compared to other methods concerning the TLI index. This suggests that the model fit worse based on GLS when the models were correctly specified. As the level of response categories increased, GLS was able to improve its TLI performance, but still below a satisfactory level. Again, both DWLS and ULS were able to maintain their consistency concerning TLI and CFI indices above 0.9 across different experimental conditions.

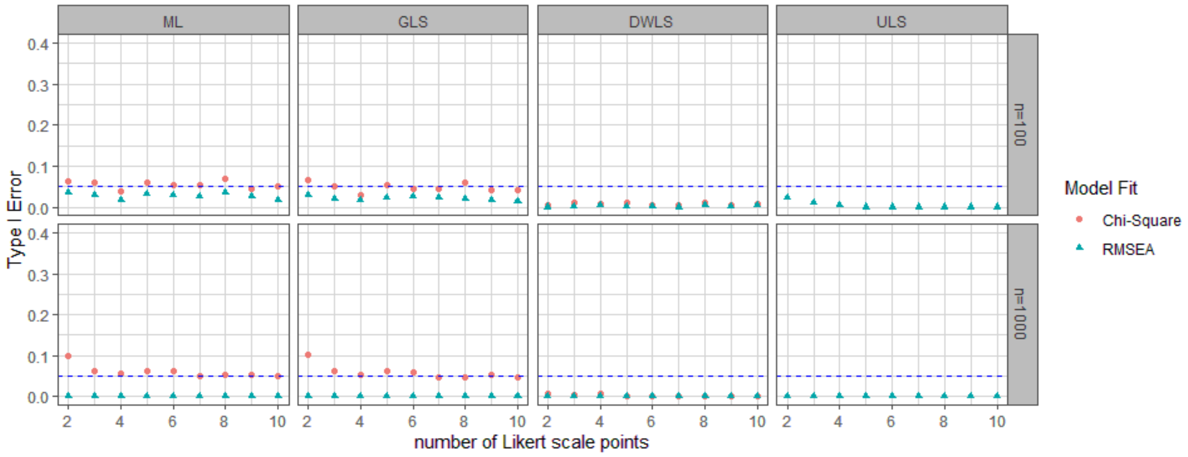

Figure 4. Plot of rejection rate from Chi-Square test and RMSEA for a one-factor model with 4 indicators per factor.

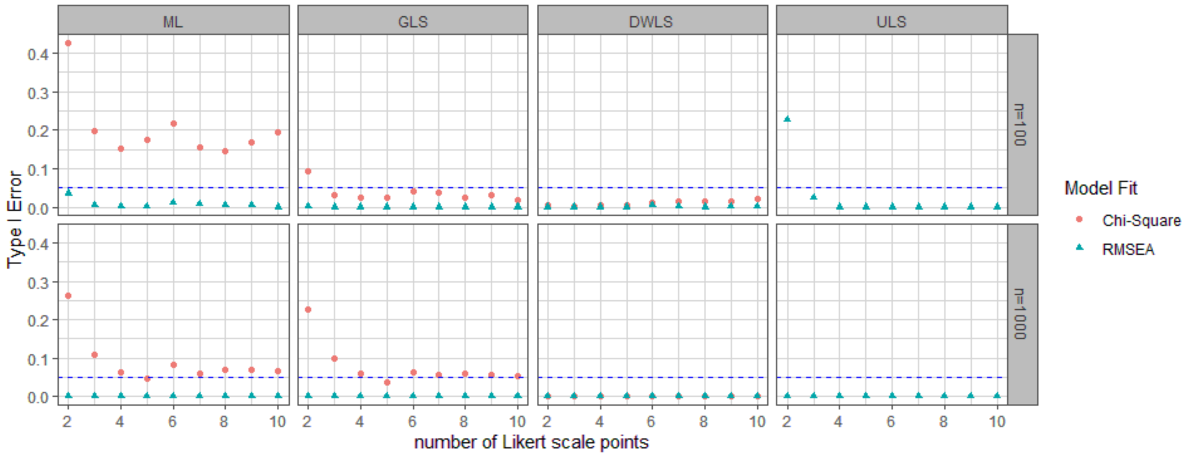

Figure 5. Plot of rejection rate from Chi-Square test and RMSEA for a two-factor model with 8 indicators per factor.

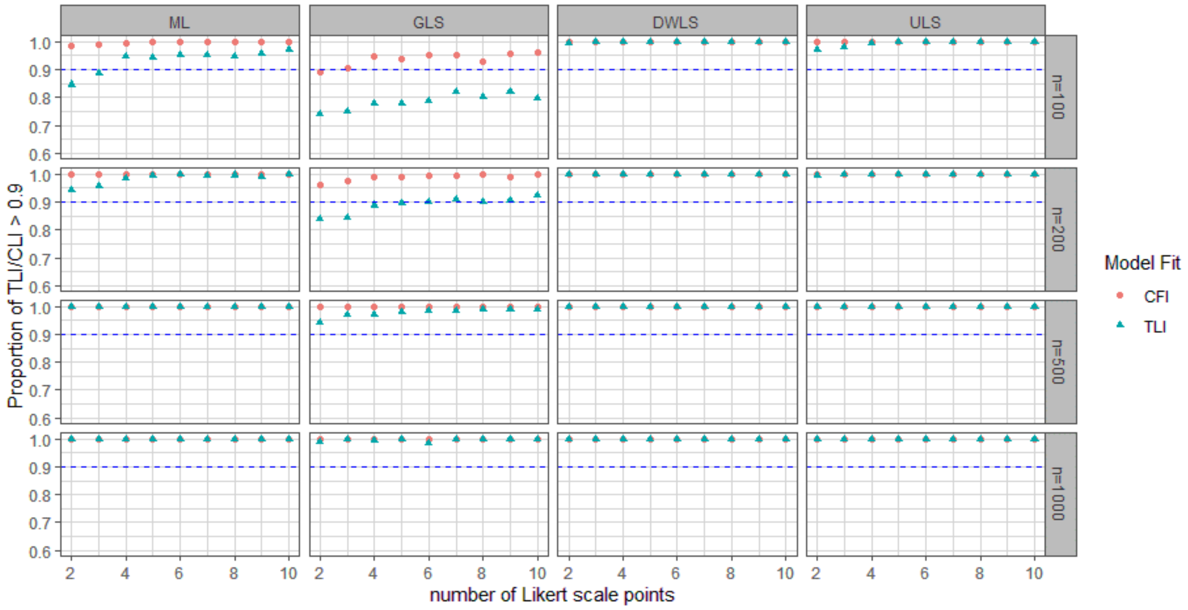

Figure 6. Plot of a proportion of TLI/CLI values above 0.9 for a one-factor model with 4 indicators per factor.

Empirical Studies

To support the findings from the simulation studies, empirical studies were carried out using available data from open-source psychometrics projects (https://openpsychometrics.org/rawdata/) [32]. Three datasets were considered. The first dataset was collected from the Taylor Manifest Anxiety Scale (TMAS). It consisted of 50 true-false statements to identify the eligibility criteria for individuals for enrolling in the stress studies and other related psychological phenomena [33]. The second dataset was obtained from the Nature Relatedness Scale (NR-6) to measure how an individual's trait level of feeling emotionally connected to nature [34]. The test consisted of six statements of opinions rated on a five-point scale of how many people agree with each of the statements. The third dataset was collected from the Right-wing Authoritarianism Scale (RWAS) to understand the psychologies of fascist regimes and their followers [35] [36]. A total of 22 statements of opinions was used in the test rated on a nine-point scale from very strongly disagree (1) to very strongly agree (9). Not all observations in the datasets are used in building the CFA models, but instead, the samples were randomly selected from the main datasets, thus each study using 750-1500 observations.

Table 1 summarizes model fit indices from fitting a one-factor model to the three studies. The Chi-square tests were not able to detect a good fit model across all studies. This can be expected since the test is highly sensitive to sample size. Alternatively, RMSEA statistics can be used to assess model fit. In the TMAS study, the estimated RMSEA value based on ULS was above 0.05 or associated with a p-value < 0.001, indicating a poor fit. This is in line with the findings in Figure 5, that ULS performed the worst when two response categories are used. Concerning CFI and TLI indices, both ML and GLS were not able to reach satisfactory fit since both values were below 0.9, particularly the TLI index obtained from the GLS method showed a strong underestimation. This is in agreement with the results shown in Figure 6. DWLS outperformed the other methods in the study with two Likert scale points. Meanwhile, in the NR-6 and RWAS studies, both DWLS and ULS performed almost equally well. In the case of the RWAS study, only ULS indicated a good fit based on RMSEA criteria. The associated factor loading estimates are reported in Table 2. All loadings were statistically significant at a 5% level, irrespective of the estimation methods used. However, all coefficients based on ML and GLS were consistently lower than DWLS and ULS.

Table 1. Summary of model fit indices for TMAS, NR-6, and RWAS studies (cases in boldface indicated a good model fit)

Study	Estimation Method	Chi-square Test			RMSEA		CFI	TLI
		statistic	df	p-value	statistic	p-value		
TMAS (50 indicators, 2-point scale)	ML	14494.4	1175	<0.001	0.050	**0.203**	0.718	0.706
	GLS	8369.3	1175	<0.001	0.037	**1.000**	0.210	0.176
	DWLS	13559.6	1175	<0.001	0.049	**0.999**	**0.945**	**0.943**
	ULS	-	-	-	0.070	<0.001	**0.952**	**0.950**
NR-6 (6 indicators, 5-point scale)	ML	133.5	9	<0.001	0.095	<0.001	**0.964**	**0.939**
	GLS	117.9	9	<0.001	0.089	<0.001	0.842	0.737
	DWLS	62.2	9	<0.001	0.062	**0.074**	**0.996**	**0.993**
	ULS	-	-	-	0.043	**0.183**	**0.993**	**0.989**
RWAS (22 indicators, 9-point scale)	ML	1923.6	209	<0.001	0.105	<0.001	0.855	0.840
	GLS	1017.0	209	<0.001	0.072	<0.001	0.289	0.214
	DWLS	1361.8	209	<0.001	0.087	<0.001	**0.992**	**0.991**
	ULS	-	-	-	0.050	**0.507**	**0.994**	**0.993**

Table 2. Estimated standardized factor loadings for TMAS, NR-6, and RWAS studies (only the first five-factor loadings are presented)

Study	Indicator	ML	GLS	DWLS	ULS
TMAS	TMAS1	0.320	0.334	0.417	0.419
	TMAS2	0.324	0.353	0.451	0.444
	TMAS3	0.258	0.266	0.337	0.344
	TMAS4	0.339	0.353	0.450	0.449
	TMAS5	0.277	0.285	0.358	0.361
NR-6	NR-61	0.503	0.527	0.557	0.560
	NR-62	0.540	0.546	0.586	0.588
	NR-63	0.779	0.793	0.837	0.805
	NR-64	0.598	0.622	0.726	0.733
	NR-65	0.848	0.846	0.891	0.899
RWAS	RWAS1	0.559	0.660	0.638	0.612
	RWAS2	0.715	0.792	0.798	0.791
	RWAS3	0.806	0.857	0.859	0.834
	RWAS4	0.757	0.811	0.831	0.820
	RWAS5	0.672	0.771	0.737	0.715

Discussion

The comparison of CFA model performance using different estimation methods (ML, GLS, DWLS, and ULS) was discussed in this study under various experimental conditions such as factor dimensionality, number of indicators, number of Likert scale points, and sample size. It is worth highlighting that each method has its strengths and shortcomings. This study gives clear evidence that DWLS and ULS outperform the ML method in all conditions. Meanwhile, GLS presents the worst performance, particularly related TLI index to measure model fit. ULS and DWLS consistently provide accurate factor loading estimates even in smaller sample sizes.

This study also reveals that using at least a 5-point Likert scale, the data can be treated as continuous and use the ML method to estimate the model parameters since the resulting parameter values are found to have trivial bias. However, the drawback is that it is very unlikely to achieve a satisfactory fit based on the Chi-square test, and it is more profound when using more indicators in a two-factor model. This, of course, has implications in empirical studies. The estimated parameters can only be interpreted and be trusted once the model shows no lack of fit.

The choice of estimators (ML, DWLS, and ULS) when evaluating model fit based on RMSEA do not seem to have an impact on the rate of rejection, since all methods are able to control the type I error at a 5% level. With an exception of ULS, which performs poorly under small sample size and two-response categories. The result from the empirical study also supports these findings. The model fit based on TLI and CFI performs almost equally well, except ML that only works best when the number of response categories increases. The TLI index is rather conservative under GLS and it is very unlikely to achieve a satisfactory fit as compared to CFI, particularly with 3 or fewer response categories in the condition of a small sample size.

It is important to note that any working recommendations provided in this study are based on the current model configurations. This research did not take into account the possible violation such as non-normality in latent distribution and asymmetric threshold. These types of violations would be interesting to investigate and see how the impact of different estimation methods on model performance. A future study should also address the estimators' behavior when the CFA model is not correctly specified, such as omitting important factor loadings and ignoring significant inter-correlation between two or more factors.

CONCLUSIONS

The present findings suggest the importance to select appropriate estimation methods based on how the data are measured. The study focuses on a type of data collected using a survey instrument measured in a Likert-type scale, which is ordinal in nature. Maximum likelihood as the most common estimation method used, which has several nice properties, failed to give accurate parameter estimation when the data are ordinal, particularly with fewer numbers of Likert scale points. In addition, the goodness of fit test, particularly the Chi-square test, under ML generally gives evidence that the model fits badly given the fact that the model is correctly specified. GLS is clearly not recommended to be used in CFA with ordinal data. The simulation results enable us to deliver clear suggestions to applied researchers when dealing with ordinal data by using the robust categorical methodology, such as DWLS and ULS. Both methods yield consistently accurate parameter estimation regardless of the number of Likert scale points and sample size. However, it should be noted that DWLS is preferable when dealing with two response categories.

ACKNOWLEDGMENTS

This research is financially supported by UIN Syarif Hidayatullah Jakarta with BLU research grant scheme no. UN.01/KPA/760/2021.

REFERENCES

[1] B. D. Zumbo, Validity: Foundational Issues and Statistical Methodology. In C. R. Rao & S. Sinharay (eds.), Handbook of statistics, Vol. 26: Psychometrics(pp. 45-79), Amsterdam: Elsevier Science, 2007.

[2] F. P. Holgado-Tello , M. A. Morata-Ramírez and Barbero-Garcí, "Confirmatory Factor Analysis of Ordinal Variables: A Simulation Study Comparing the Main Estimation Methods," *Avances en Psicología Latinoamericana,* vol. 36, no. 3, pp. 601-617, 2018.

[3] B. Thompson, Explanatory and Confirmatory Factor Analysis., Washington DC: American Psychological Association, 2004.

[4] K. G. Joreskog and D. Sorbom , Lisrel 8, User's Reference Guide., Chicago: SSI Inc. (Scientific Software International), 1996.

[5] Y. Li , Confirmatory Factor Analysis with Continuous and Ordinal Data: An Empirical Study of Stress Level, Uppsala University, 2014.

[6] T. A. Brown, Confirmatory factor analysis for applied research, Guilford Press, 2012.

[7] K. A. Bollen, Structural equations with latent variables, New York: Wiley, 1989.

[8] J. F. Finch , S. G. West and D. P. MacKinnon , "Effects of sample size and nonormality on the estimation of mediated effects in latent variable models," *Structural Equation Modeling,* vol. 4, pp. 87-107, 1997.

[9] B. O. Muthén and D. Kaplan , "A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model," *British Journal of Mathematical and Statistical Psychology,* vol. 45, pp. 19-30, 1992.

[10] S. G. West, J. F. Finch and P. J. Curran, Structural equation models with nonnormal variables: problems and remedies. In R. H. Hoyle (Ed.), Structural Equation Modeling: concepts, issues and applications (pp. 56-75)., Thousand Oaks (CA): Sage, 1995.

[11] A. Maydeu-Olivares , "Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling," *Psychometrika,* vol. 66, pp. 209-227, 2001.

[12] R. Tate, "A comparison of selected empirical methods for assessing the structure of responses to test items," *Applied Psychological Measurement,* vol. 27, no. 3, p. 159–203, 2003.

[13] C. G. Forero, A. Maydeu-Olivares and D. Gallardo-Pujol , "Factor analysis with ordinal indicators: a monte carlo study comparing DWLS and ULS simulation," *Structural Equation Modeling,* vol. 16, pp. 625-641, 2009.

[14] D. Shi and A. Maydeu-Olivares , "The effect of estimation methods on SEM fit indicies," *Educational and Psychological Measurement,* vol. 80, no. 3, pp. 421-445, 2020.

[15] T. P. Morris , "White IR, Crowther MJ. Using simulation studies to evaluate statistical methods," *Statistics in Medicine,* vol. 38, pp. 2074-2102, 2019.

[16] A. Satorra, "Robustness issues in structural equation modeling: A review of recent developments," *Quality and Quantity,* vol. 24, no. 4, p. 367–386, 1990.

[17] M. W. Browne, "Generalized least-squares estimators in the analysis of covariance structures.," *South African Statistical Journal,* vol. 8, pp. 1-24, 1974.

[18] L. Ding, W. Velicer and L. Harlow , "Effect of estimation methods, number of indicators per factor and improper solutions on structural equation modeling fit indices," *Structural Equation Modeling,* vol. 2, pp. 119-143, 1995.

[19] U. H. Olsson , T. Foss , S. V. Troye and R. D. Howell , "The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality," *Structural equation modeling,* vol. 7, no. 4, pp. 557-595, 2000.

[20] W. F. Yang , K. G. Jöreskog and H. Luo, "Confirmatory factor analysis of ordinal variables with misspecified models," *Structural Equation Modeling,* vol. 17, no. 3, p. 392–423, 2010.

[21] B. Muthén, S. H. C. du Toit and D. Spisic , Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes, Unpublished manuscript, 1997.

[22] H. W. Marsh, K. Hau , J. R. Balla and D. Grayson, "Is more ever too much? The number of indicators per factor in confirmatory factor analysis," *Multivariate Behavioral Research,* vol. 33, p. 181–220, 1998.

[23] P. Paxton , P. J. Curran, K. A. Bollen , J. Kirby and F. Chen, "Monte Carlo experiments: Design and implementation," *Structural Equation Modeling A Multidisciplinary Journal,* vol. 8, no. 2, pp. 287-312, 2001.

[24] C. DiStefano, "The impact of categorization with confirmatory factor analysis," *Structural Equation Modeling,* vol. 9, p. 327–346, 2020.

[25] C. -H. Li , "Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares," *Behav Res,* vol. 48, pp. 936-949, 2016.

[26] L. K. Muthén and B. O. Muthén, "How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power," *Structural Equation Modeling A Multidisciplinary Journal ,* vol. 9, no. 4. , 2009. DOI: 10.1207/S15328007SEM0904_8..

[27] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software,* vol. 48, no. 2, pp. 1-36, 2011.

[28] The R Development Core Team, R: A language and environment for statistical computing., Vienna, Austria: : R Foundation for Statistical Computing, 2019.

[29] M. Wallentin, A. H. Nielsen and M. Friis-Olivarius, "The Musical Ear Test, a new reliable test for measuring musical competence," *Learning and Individual Differences,* vol. 20, pp. 188-196, 2010.

[30] D. Kaplan, "A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models," *Multivariate Behavioral Research,* vol. 24, p. 41–57, 1989.

[31] P. J. Curran , S. G. West and G. F. Finch , "The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis," *Psychological Methods,* vol. 1, p. 16–29, 1996.

[32] Open-Source Psychometrics Project, "Raw data from online personality tests," https://openpsychometrics.org/_rawdata/, 2019.

[33] J. A. Taylor, "A personality scale of manifest anxiety," *The Journal of Abnormal and Social Psychology,* vol. 48, no. 2, pp. 285-290, 1953.

[34] E. Nisbet and J. M. Zelenskix, "The NR-6: A new brief measure of nature relatedness," *Frontiers in Psychology,* vol. 4, 2013. DOI: 10.3389/fpsyg.2013.00813.

[35] B. Altemeyer, Right-wing authoritarianism, University of Manitoba Press, 1981.

[36] B. Altemeyer, The Authoritarians, University of Manitoba, 2007.