International Journal of Assessment Tools in Education
2021, Vol. 8, No. 1, 21-37

https://doi.org/10.21449/ijate.782351

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Comparison of confirmatory factor analysis estimation methods on mixed-format data

Abdullah Faruk Kilic [1,*], **Nuri Dogan** [2]

[1]Adıyaman University, Faculty of Education, Department of Educational Sciences, Turkey
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Turkey

ARTICLE HISTORY

Received: Aug. 27, 2020
Revised: Nov. 16, 2020
Accepted: Jan. 05, 2021

Keywords:

Bayesian estimation,
Monte-carlo simulations,
CFA.

Abstract: Weighted least squares (WLS), weighted least squares mean-and-variance-adjusted (WLSMV), unweighted least squares mean-and-variance-adjusted (ULSMV), maximum likelihood (ML), robust maximum likelihood (MLR) and Bayesian estimation methods were compared in mixed item response type data via Monte Carlo simulation. The percentage of polytomous items, distribution of polytomous items, categories of polytomous items, average factor loading, sample size and test length conditions were manipulated. ULSMV and WLSMV were found to be the more accurate methods under all simulation conditions. All methods except WLS had acceptable relative bias and relative standard error bias. No method gives accurate results with small sample sizes and low factor loading, however, the ULSMV method can be recommended to researchers because it gives more appropriate results in all conditions.

1. INTRODUCTION

Evidence for validity should be collected first in test development or adaptation studies.The process of collecting validity evidence for a test's structure mostly involves examining the relationships between the variables (Bollen, 1989). Factor analysis is one of the oldest and best known ways to investigate relationships between variables (Byrne, 2016; Osborne & Banjanovic, 2016). The use of confirmatory factor analysis (CFA) in the process of collecting evidence of construct validity is an accepted approach in the literature, and thus frequently used (AERA et al., 2014; DiStefano & Hess, 2005; Guilford, 1946; Nunnally & Bernstein, 1994; Thompson & Daniel, 1996). A search for the key term "confirmatory factor analysis" in the Scopus database resulted in 34.257 articles. When the search was limited to the field of psychology and social sciences, there were 19.546 articles. 461 of these articles were published in 2020. Confirmatory factor analysis is thus frequently used in the field of social sciences and psychology.

The use of CFA requires knowledge of which estimation method provides accurate results under which conditions, because estimation methods affect the results obtained when estimates

CONTACT: Abdullah Faruk KILIC ✉ abdullahfarukkilic@gmail.com Adıyaman University, Faculty of Education, Deparment of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

were biased. There are thus numerous studies in the literature comparing CFA estimation methods. An examination of studies in which the observed variables were categorical found that some studies were performed with only five categories of observed data. The manipulated simulation conditions were the distribution of the observed or latent variables, the estimation methods used and the sample sizes in these studies (Babakus et al., 1987; B. O. Muthén & Kaplan, 1985, 1992; DiStefano, 2002; Ferguson & Rigdon, 1991; Lei, 2009; Morata-Ramirez & Holgado-Tello, 2013; Potthast, 1993). Examining other simulation studies with categorical data found that there were between two and seven categories of observed variables (Beauducel & Herzberg, 2006; Dolan, 1994; Flora & Curran, 2004; Green et al., 1997; Li, 2016; Liang & Yang, 2014; Moshagen & Musch, 2014; Rhemtulla et al., 2012; Yang-Wallentin et al., 2010). Studies comparing estimation methods on mixed item response type data, however, were few and limited (Depaoli & Scott, 2015; Oranje, 2003).

The study by Depaoli and Scott (2015) was retracted due to systematic error in the simulation codes. Item type (including different combinations of item types), factor loadings, factor correlations, sample sizes, and priors in the case of Bayesian conditions was examined, however, the percentage and distributions of polytomous items were not manipulated.In the simulation study conducted by Oranje (2003), sample size, number of factors, number of observed variables per factor, and item response-type were manipulated, and ML, WLS and WLS (estimated to Lisrel software), WLSM and WLSMV (estimated to Mplus software) estimation methods were compared. The study reported that as the number of categories increases, the sensitivity of the parameter estimates increases, because polychoric correlations are more appropriate in this condition. However, the distribution of polytomous items was not manipulated in this study, and the study was conducted in a single mixed format test (60% with 2 categories, 20% with 3 categories and 20% with 5 categories).

1.1. The Present Study

Despite the large number of studies comparing CFA estimation methods, there does not seem to be a study comparing both frequentist and Bayesian estimation methods in terms of mixed item response type data. Therefore, investigating this comparison will close this gap in the CFA literature. In addition, the current study studied in a large number of simulation conditions to close this gap. So, the current study can meet the needs of applied researchers who use CFA to collect validty evidence. This study will thus contribute to the literature on CFA estimation methods.

This study investigates which CFA estimation method gives unbiased and accurate results for simulation conditions with mixed item response type data. Research problems were therefore constructed as follows. According to the simulation conditions; which estimation methods have more accurate i) convergence rate and inadmissible solution rate, ii) percentage of accurate estimate (PAE), iii) relative bias (RB), iv) standard error bias (SEB) values? and v) how accurate is the performance of ML, MLR, ULSMV, WLS, WLSMV and Bayesian on four different empirical data sets in terms of convergence, inadmissible solution rate, RB and SEB values?

2. METHOD

A Monte Carlo simulation was used in the present study. Monte Carlo studies are statistical sampling investigations. In these studies, dataset suitable for empirical distribution is generated. The aim of these studies is to produce a data set suitable for empirical distribution. This situation separates Monte Carlo studies from simulation studies. Because in simulation studies, it is possible to generate dataset for population or to demonstrate a statistical analysis. However, sample data are generated in accordance with a certain distribution in Monte Carlo simulations (Bandalos & Leite, 2013). It compared CFA estimation methods in mixed item response type data. Simulation and empirical data sets were both used in the study. The empirical data set included four tests of the Monitoring and Evaluation of Academic Skills (MEAS) research data sets which conducted by Turkish Ministry of National Education. The tests consist of 18 items, some items are scored as 1-0, some items are 0-1-2 and some 0-1-2-3. The tests included both binary and polytomous items.

2.1. Manipulated Factors

This study focused on achievement tests consisting of mixed item responses. Mixed item response type achievement tests are generally reported to be unidimensional (Bennett et al. 1990; Bennett, Rock, and Wang 1991; Lissitz, Hou, and Slater 2012; van den Bergh 1990). For this reason, the measurement model was defined as unidimensional. The percentage of polytomous items ((10%, 20%, 40%, 50%), skewness of polytomous items (left skewed, normal, right skewed), categories of polytomous items (3, 4 and 5), average factor loading (.40, .60 and .80), sample size (200, 500 and 1000) and test length (20, 30 and 40 items) were manipulated as simulation conditions. The simulation conditions were fully crossed, so, 972 (4x3x3x3x3x3) simulation conditions were manipulated, with 1000 replicates per cell.

The average factor loading was chosen as low (.40), medium (.60) and high (.80). Since the lowest factor loading in such tests is recommended as .40 (Howard, 2016), the low value of the average factor loading is.40, medium is.60 and high is.80. It is not common in practice that all items have the same factor loading, and so unlike other studies, the factor loadings of all the items in the test were not equal (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Forero et al., 2009; Li, 2016a; Liang & Yang, 2014).

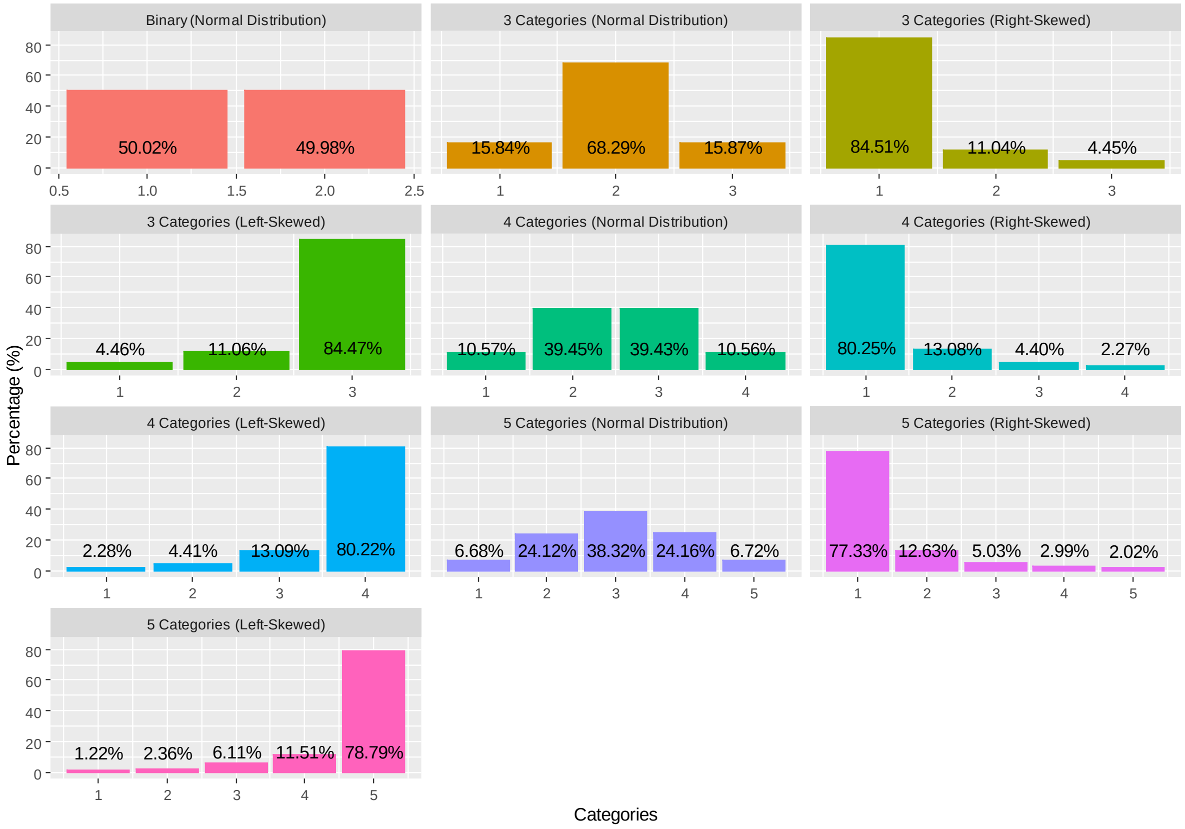
Sample sizes were determined as 200 (small), 500 (medium) and 1000 (large), as used in many other simulation studies (Beauducel & Herzberg, 2006; Li, 2016a; Oranje, 2003; West et al., 1995).

Considering the real test situations, the percentage of polytomous items and the categories of polytomous items were determined as 10%, 20%, 40%, 50%, and 3, 4, 5 respectively. Since it is thought that the distribution of polytomous items may have an impact on the estimates, the distribution of polytomous items was added to the simulation conditions as left-skewed, normal and right-skewed. The test length was manipulated to be short (20 items), medium (30 items) and long (40 items).

2.2. Data Generation

Continuous data sets (continuous latent variable) were first generated for each condition of the study, followed by multivariate normal distribution. Once the continuous data sets were generated, the data was categorized according to simulation conditions. This approach is commonly used in the literature (Beauducel & Herzberg, 2006; Lei, 2009; Morata-Ramirez & Holgado-Tello, 2013; Oranje, 2003; T. K. Lee et al., 2018). This approach also meets the assumption that the underlying variable is normally distributed in psychology (Crocker & Algina, 2008; Gulliksen, 1950). Continuous data sets were categorized as binary (normally distributed), 3 categories (left-skewed, normal and right-skewed), 4 categories (left-skewed, normal and right-skewed) and 5 categories (left-skewed, normal and right-skewed). The distribution of categorical variables used in the study is presented in in Figure 1.

Figure 1. *Distribution of variables.*



2.3. Outcome Variables

Non-convergence or inadmissible solutions rate, relative bias for factor loadings (RB), percentage of accurate estimates for factor loadings (PAE) and standard errors bias (SEB) were used as outcome variables in the study.

Since 1000 there were replications in the study, estimation methods with 500 or more nonconvergence or inadmissible solutions were considered “NA” for that condition.

Relative bias was calculated via

$$RB = \frac{\hat{\varphi} - \varphi_{True}}{\varphi_{True}} \tag{1}$$

where $\hat{\varphi}$ is the mean of sample estimates over the 1000 replications of average factor loading and φ_{True} is the true average factor loading. In the literature, |RB|<.05 indicates trivial bias, .05 ≤|RB|≤ .10 indicates moderate bias and |RB|> 0.10 indicates substantial bias (Flora & Curran, 2004; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Moshagen & Musch, 2014; Rhemtulla et al., 2012). |RB| ≤0.10 was thus considered “acceptable” in this study.

To determine the percentage of accurate estimate (PAE), the average factor loading value obtained from 1000 replications was examined as to whether it was within ± 5% of the real factor loading determined in the simulation condition (Ferguson & Rigdon, 1991; Wolf, Harrington, Clark, & Miller, 2013). Methods with 95% or more of PAE were considered “acceptable” in this study.

Standard error bias (SEB) was calculated via

$$SEB = \frac{\frac{1}{n_{rep}} \Sigma_{t=1}^{n_{rep}} \widehat{se}(\hat{\theta}_{pt})}{sd(\hat{\theta}_{pt})} \tag{2}$$

where $\widehat{se}(\hat{\theta}_{pt})$ was the standard error of parameter p for replication t, $sd(\hat{\theta}_{pt})$, was the standard deviation of parameter estimates obtained from t replications (Forero et al., 2009; Holtmann, Koch, Lochner, & Eid, 2016; Rhemtulla et al., 2012). When the standard error estimates are equal to the standard deviation obtained empirically, the SEB value will be equal to 1. Accordingly, the SEB value was classified as follows (Holtmann et al., 2016): 5/6 <SEB<6/5 was negligible, 2/3 <SEB<5/6 and 6/5 <SEB<3/2 was medium and SEB<2/3 or SEB> 3/2 was large.

The Psych Package (Revelle, 2019) in the R software (R Core Team, 2018) was used to generate the simulation data. Mplus software (L. K. Muthén & Muthén, 2012) was used for CFA. Since 1000 replications were used in the study, the data sets were analyzed in Mplus software using the MplusAutomation (Hallquist & Wiley, 2017) package.

2.4. Data Analysis in Real Data Sets

The empirical data sets were obtained from the Monitoring and Evaluation of Academic Skills (MEAS) research carried out in 2016 in Turkey. Different item types were used in the MEAS research. For this reason, "Open-Ended and Multiple-Choice Question Writing" training was given to item writers by academicians. It is emphasized that the prepared items were reviewed by measurement and evaluation experts and language experts, and after the necessary arrangements, a pilot application was carried out with approximately 5000 students in Ankara, Turkey. The actual application of MEAS research was conducted with the participation of about 38.000 students from 81 provinces of Turkey (MoNE, 2017). A rubric was developed for scoring open-ended items. Accordingly, firstly, correct and partially correct answers were formed. After the pilot application, the answers given by the students to the open-ended items were examined and the unpredictable answers were added to the rubric. Thus, the rubric was composed of four parts as true, partial true, false and empty. The research was conducted n the fields of Turkish, mathematics, science and social studies. The reliability coefficient of the tests as internal consistency ranged between .73 to .85. Test data from Turkish which was focused on reading comprehension (13 binary, 5 three categories), mathematics (12 binary, 6 three categories), science (14 binary, 4 three categories) and social sciences (15 binary, 2 three categories and 1 five categories) was used. Missing data was removed from the data sets via listwise deletion. After removal, the Turkish, mathematics, science and social studies test data consisted of 4745, 2247, 3143 and 3442 individuals, respectively.

Sampling was first undertaken for each data set. Since the sample size conditions were determined as 200, 500 and 1000, the same sample sizes were randomly taken from the Turkish focused on reading comprehension, mathematics, science and social studies test data sets. Sampling was repeated 100 times for each test, in order to avoid the sample bias.

The outcome variables in the analysis performed with real datasets were non-convergence or inadmissible solutions rate, relative bias for factor loadings (RB), and percentage of accurate estimates for factor loadings (PAE).

The true parameter value was needed to calculate the PAE and RB value. The true average factor loading value of the real data sets is unknown. The true value of the average factor loading was obtained using exploratory factor analysis (EFA). For this purpose, AFA was conducted with the whole sample in the Turkish, mathematics, science and social studies datasets. Unweighted least squares (unweighted least square [ULS]), which is claimed to be strong against the assumption that multivariate normality is severely violated, was used as a factor extraction method (Nunnally & Bernstein, 1994; Osborne & Banjanovic, 2016). EFA demonstrated that the data sets were unidimensional.

A repeated measures ANOVA was used to determine which simulation factor is more effective on PAE, RB and r-SEB values. Since the same data sets were analyzed using different estimation methods, the estimation methods are defined as within-subject. The simulation conditions are defined as between-subject. Partial η^2 was used to examine the effect size. In partial η^2 .01 or less is interpreted as being a small, .06 or more a medium and .14 or more a large effect (Cohen, 1988).

In the real data set, analyses for EFA were performed using Factor 10.08 software (Lorenzo-Seva & Ferrando, 2020).

3. RESULT / FINDINGS

3.1. Convergence and Inadmissible Solution Rate

The convergence rates of maximum likelihood (ML), robust maximum likelihood (MLR) and Bayes are 100% and their inadmissible solutions rates are 0%. A detailed table for the convergence and inadmissible solution rates of other methods is given in Appendices A-E.

The convergence rate of the unweighted least squares mean-and-variance-adjusted (ULSMV) method is 100% and the inadmissible solution rate is 0.01%. The ULSMV method has an inadmissible solution under conditions where the sample size is 200, skewed 3 or 4 category polytomous items, and average factor loading is .80. The coverage rate of all methods except Bayesian is over 90% for all models.

The convergence rate of the weighted least squares mean-and-variance-adjusted (WLSMV) method was 99.99%, while its inadmissible solution rate was .02%. Data sets seem to have convergence problems under conditions where the sample size is 200, the average factor loading is .40, and the test length is 30 or 40 items for WLSMV method. There are inadmissible solutions in conditions similar to ULSMV where the sample size is 200, polytomous items were skewed, the average factor loading was .80, and there were polytomous items in 3 or 4 categories.

The convergence rate of the weighted least squares (WLS) method is 49.48%, and the inadmissible solution rate is 7.03%. The WLS method was not converged under any conditions with a sample size of 200. Additionally, when the sample size was 500, it was not converged under any conditions where the test length was 40 items. Accordingly, it can be said that the WLS method does not converge in small samples or long tests.

There was convergence problem for the WLS method when increasing the number of polytomous items under conditions where sample size was 500, test length was 30 items and average factor loadings were .40 and .60. Increasing the number of polytomous items categories to five resulted in convergence problems under conditions where percentage of polytomous items were 40% and %50. The convergence problems of the WLS method decreased as the sample size increased to 1000.

Examination of the inadmissible solutions in the WLS method suggests that this method has more inadmissible solutions under conditions where the sample size was 1000 and the average factor loading was .80. WLS has inadmissible solutions in about 40% of all data sets under conditions where sample size was 500, test length was 30 items, and the average factor loading was .60.

3.2. Percentage of Accurate Estimates

The PAE of WLS method was not examined due to its low convergence rate. The PAE values of other methods are presented in Appendix F in detail, for all conditions.

Under conditions where the sample size was 200 and average factor loadings were .40 and .60, the PAEs of the all estimation methods were less than 95%. When increasing the average factor loading to .80, the PAE of the methods were greater than 95%. Under 36 conditions where sample size was 200, the average factor loading was .80, and polytomous items had 3 categories

(3 conditions of distributions of polytomous items x 3 conditions of test length x 4 conditions of percentage of polytomous items = 36 conditions), the Bayesian method's PAE values were greater than 95% in more conditions (33 conditions). For the specified simulation conditions (3 conditions of distributions of polytomous items x 3 conditions of test length x 4 conditions of percentage of polytomous items = 36 conditions), the PAE values of the ULSMV method were close to those of the Bayesian method (26 conditions). Under conditions where sample size was 200 and 3 categories of polytomous items followed normal distribution, WLSMV, ULSMV and Bayesian methods had similar PAE values, but the distribution of polytomous items was skewed, and the WLSMV method's PAE values decreased. Under conditions where the sample size was 200, polytomous items had 4 or 5 categories, and polytomous items followed normal distribution, the PAE values were bigger n the Bayesian and ULSMV methods than the ML/MLR and WLSMV methods. When the ML/MLR, and WLSMV methods were compared, the WLSMV method had bigger PAE values.

The PAE value of the methods was below 95% in all conditions with a sample size of 500 and an average factor loading of .40. When the average factor loading increased to .60 and .80, the PAE value of the methods increased to 95%. When the sample size increased to 1000, the PAE values of ULSMV, WLSMV and Bayesian methods exceeded 95% under some conditions with an average factor loading of .40. Accordingly, it can be said that the PAE values increase in the estimation methods when sample size or average factor loading increase.

A repeated measures ANOVA was performed to determine which simulation condition was more effective as regards PAE values. In Mauchly's Test of Sphericity, sphericity was violated ($\chi^2(5)$ = .01, p <.001). The Greenhouse-Geisser correction was thus used. There was a statistically significant main effect of the estimation method on PAE values overall F(1.53, 1357.14) = 27797.19, p = .00, partial η^2 = .97).

When the average PAE values of the methods were compared with the Bonferroni correction, ULSMV (mean = 89.56%, se = .55) was statistically significantly higher than other methods. The WLSMV (mean = 89.17%, se = .56) method's PAE was statistically significantly higher than both Bayesian (mean = 87.56%, se = .55) and ML/MLR (mean = 67.77%, se = 1.07) methods. The Bayesian method's PAE value, on the other hand, was statistically significantly higher than in the ML/MLR method.

When the test of within-subject effects was examined, the most important second order interaction was found to be method x average factor loading (F(3.06, 1357.14) = 9053.40, p = .00, partial η^2=.95). The other second order interactions method x sample size (F(3.06, 1357.14) = 1705.99, p = .00, partial η^2=.79), method x percentage of polytomous item (F(3.06, 1357.14) = 877.43, p = .00, partial η^2=.75), method x distribution of polytomous item (F(3.06, 1357.14) = 294.42, p = .00, partial η^2=.40) and method x categories of polytomous items (F(3.06, 1357.14) = 112.99, p = .00, partial η^2=.20) had a large effect size, but the interaction of method x test length (F(3.06, 1357.14) = 14.57, p = .00, partial η^2=.03) had a small effect size.

When the third order interactions were examined, the most important third order interaction was found to be method x average factor loading x sample size (F(6.11, 1357.14) = 1106.00, p = .00, partial η^2=.83). The other third order interactions method x sample size x percentage of polytomous items (F(9.17, 1357.14) = 224.91, p = .00, partial η^2=.60), method x distribution of polytomous items x average factor loading (F(6.11, 1357.14) = 61.04, p = .00, partial η^2=.22), method x percentage of polytomous items x sample size (F(9.17, 1357.14) = 40.66, p = .00, partial η^2=.22), method x average factor loading x test length (F(6.11, 1357.14) = 46.66, p = .00, partial η^2=.17) and method x categories of polytomous items x average factor loading (F(6.11, 1357.14) = 36.14, p = .00, partial η^2=.20) had a large effect size. The other interactions had medium and small effect sizes ranging between .01-.13.

Between-subject effect was examined to investigate which simulation condition had a higher effect on PAE values. Average factor loading had the biggest effect on the PAE values (F(2, 888) = 41569.78, *p* = .00, partial η^2 = .99). Sample size (F(2, 888) = 10670.12, *p* = .00, partial η^2 = .96), percentage of polytomous items (F(3, 888) = 376.89, *p* = .00, partial η^2 =. 56), test length (F(2, 888) = 356.13, *p* = .00, partial η2 = .45) and categories of polytomous items (F(2, 888) = 6.20, *p* = .00, partial η^2 =. 01) had an effect on the PAE values, however, PAE values do not differ significantly according to the distribution of polytomous items (F(2, 888) = 0.49, *p* = .62, partial η^2 <.00).

Because there were many between subject variables, only second and third order interactions were studied. When second order interactions were examined, the important interaction was found to be average factor loading x sample size (F(4, 888) = 1693.53, *p* = .00, partial η^2 = .88). When results were examined in terms of partial eta squared, the average factor loading x percentage of polytomous items (F(6, 888) = 79.31, *p* = .00, partial η^2 = .35), and average factor loading x test length (F(4, 888) = 79.31, *p* = .00, partial η^2 = .26) interactions had large effect size. Distribution of polytomous items x sample size (F(4, 888) = 29.51, *p* = .00, partial η^2 = .12), percentage of polytomous items x sample size (F(6, 888) = 17.03, *p* = .00, partial η^2 = .10) and test length x sample size (F(4, 888) = 21.67, *p* = .00, partial η^2 = .09) had a medium effect on PAE values. The other interaction effect sizes ranged between .01-.04, and some was not statistically significant.

Examination of the post-hoc tests found that average factor loading categories differed statistically significantly from each other. So, .80 had higher PAE values than .40 and .60. Similarly, .60 had higher PAE values than .40. At the same time, sample size categories were statistically significantly different from each other: 1000 had higher PAE values than 200 and 500. Similarly, 500 had higher PAE values than 200.

Polytomous items with 3 categories had a statistically significantly higher PAE value than those with 4 and 5 categories (*p* = .01). There were no statistically significant differences between polytomous items with 4 and 5 categories. No statistically significant difference was found between the distribution of polytomous items. Accordingly, it can be said that the distribution of polytomous items has no effect on the estimation method’s PAE values.

Test length categories differed from each other statistically significantly (p = .00). So, 40 items had higher PAE values than 20 and 30. Similarly, 30 items had higher PAE values than 20. So, an increase in the number of items increases the PAE values of the methods. The percentage of polytomous items differed from each other statistically significantly (*p* = .00). So, 50% had higher PAE values than the others (10%, 20% and %40). Similarly, 40% had higher PAE values than 20% and 10%, and 20% had higher PAE values than 10%. As the percentage of polytomous items increases, therefore the PAE values of the estimation method increases.

A repeated measures ANOVA showed that the PAE values of the estimation methods differ from each other. The PAE value was obtained in the highest ULSMV method. This method was followed by WLSMV, Bayesian and ML/MLR. The most effective condition on the PAE of the methods is the average factor loading. This condition was followed by sample size (partial η^2 = .96), percentage of polytomous items (partial η^2 = .56), test length (partial η^2 = .45) and categories of polytomous items (partial η^2 = .01). When the interaction of conditions was examined, average factor loading x sample size (partial η^2 = .88) had the biggest effect on PAE values. This interaction was followed by the average factor loading x percentage of polytomous items (partial η^2 = .35), average factor loading x test length (partial η^2 = .26), distribution of polytomous items x sample size (partial η^2 = .12), percentage of polytomous items x sample size (partial η^2 = .10) and test length x sample size (partial η^2 = .09).

In summary, an increase in average factor loading, sample size, test length and percentage of polytomous items increases the PAE values of the estimation methods. Interestingly, the PAE values of the methods increased as the categories of polytomous items decreased.

3.3. Relative Bias

The RB value in the conditions converged by WLS generally decreased with an increasing number of items (substantial bias), and with a decreasing number of items, the value of RB increased (moderate bias). WLS has not been compared with other methods in which it has moderate or substantial bias under the conditions where WLS could converge. The RB values of all methods are presented in Appendix G in detail.

In the simulation conditions with a sample size of 200, the ULSMV and WLSMV had trivial RB. While the ML/MLR methods were moderately biased under conditions where average factor loading was .40, ML/MLR estimation methods have trivial RB when the average factor loading increased to .60 or .80. The Bayesian method has trivial bias in most conditions where the average factor loading was .40 and in all conditions with an average factor loading of .60 and .80.

Under conditions where the sample size was 500 and 1000, Bayesian, ULSMV and WLSMV methods had trivial bias. ML/MLR methods had trivial bias in most simulation conditions where average factor loading is .40, and in all simulation conditions where average factor loading is .60 and .80.

The RB values were acceptable (|RB| ≤ .10) for all simulation conditions in all methods except WLS. A repeated measures ANOVA was performed to examine simulation conditions affecting RB values, and thus, to examine which conditions were more effective. Mauchly's Test of Sphericity showed that sphericity was violated (χ^2 (5) = .00, *p* <.001) the Greenhouse-Geisser correction was thus used. There was a statistically significant main effect from the estimation method on RB scores overall F(1.00, 883.18) = 44.72, *p* = .00, partial η^2 = .05).

When the average RB values of the methods were compared with the Bonferroni correction, it was observed that the ULSMV (mean = -.00, se = .00) method had a statistically significantly lower RB value than other methods. The Bayesian (mean = -.01, se = .00) method is lower than both the WLSMV (mean = .02, se = .01) and ML/MLR (mean = -.04, se = .00) methods. The WLSMV method, on the other hand, has a statistically significantly lower RB value than the ML/MLR method.

When tests of within-subject effects was examined, it was observed that the most important second order interaction was method x average factor loading (F(2.00, 883.18) = 13.68, *p* = .00, partial η^2=.03) which has a small effect size. Method x sample size (F(2.00, 883.18) = 5.93, *p* = .00, partial η^2=.01) also has a small effect size, but the other second order interactions were not statistically significant.

When the third order interactions were examined, the most important third order interaction was found to be method x average factor loading x sample size (F(4, 883.18) = 4.80, *p* = .00, partial η^2=.02) which has a small effect size. Method x sample size x percentage of polytomous items (F(4, 883.18) = 2.21, *p* = .00, partial η^2=.01) also has a small effect size, but the other third order interactions were not statistically significant.

The between-subject effect was examined to investigate which simulation condition has a greater effect on RB values. Sample size had the largest effect on RB values (F(2, 883) = 3.65, *p* = .03, partial η^2 = .01). Average factor loading (F(2, 883) = 3.58, *p* = .03, partial η^2 = .01) had a smaller effect on RB values. Other simulation conditions had no effect on RB values.

When second order interactions were examined, the most important interaction was found to be test length x percentage of polytomous items (F(6, 883) = 2.23, p = .04, partial η^2 = .01) which has a small effect size. The other interactions were not statistically significant.

When post-hoc tests were examined, conditions where the average factor loading was .80 had statistically significantly smaller RB values than for .40, but there was no statistically significant difference between .80 and .60 conditions. The other simulation conditions did not affect RB values.

A repeated measures ANOVA demonstrated that the RB values of the methods differed from each other and ULSMV had the lowest RB value. This was followed by Bayesian, WLSMV and ML/MLR methods. The most effective condition regarding the RB values of the methods was sample size (partial η^2 = .01). This condition was followed by average factor loading (partial η^2 = .01). When the interaction of conditions was analyzed, method x average factor loading (partial η^2 = .03) had the largest effect on RB values.

In summary, the simulation conditions, generally, have no effect on RB values, but the condition where average factor loading was .80 had a smaller RB value.

3.4. Standard Error Bias

ML and MLR methods have negligible standard error bias in all conditions. Bayes, ULSMV and WLSMV methods were negligibly biased in most of the 200 sample size conditions. All estimation methods except WLS had negligible bias in conditions where sample size was 500 and 1000. WLS method, generally, have large bias in most conditions if converged. The SEB values obtained from the estimation methods according to the simulation conditions are presented in Appendix H for more information.

A repeated measures ANOVA was performed to examine simulation conditions affecting SEB values. Mauchly's Test of Sphericity showed that sphericity was violated ($\chi^2(9)$ = .00, p < .001) so the Greenhouse-Geisser correction was used. Estimation method had a statistically significant main effect on SEB values F(1.93, 1711.09) = 8991.97, p = .00, partial η^2=.91).

When the average SEB values of the methods were compared with the Bonferroni correction, the SEB values of the ULSMV (mean = .98 se = .00) and MLR (mean = .98 se = .00) methods differed statistically significantly from other methods and were observed to be closer to 1 (which means that there is no bias). ML (mean = .97, se = .00) differed statistically significantly from both WLSMV and Bayesian methods. The WLSMV method (mean = .96, se = .00) had a statistically significantly higher SEB value than the Bayesian method (mean = .92, se = .00).

When the test of within-subject effects was examined, the most important second order interaction was found to be method x sample size (F(3.85, 1711.09) = 2703.63, p = .00, partial η^2=.86). The other second order interactions method x average factor loading (F(3.85, 1711.09) = 572.71, p = .00, partial η^2=.56), method x categories of polytomous items (F(3.85, 1711.09) = 175.55, p = .00, partial η^2=.28), method x percentage of polytomous items (F(5.78, 1711.09) = 115.48, p = .00, partial η^2=.28), method x distribution of polytomous items (F(3.85, 1711.09) = 153.21, p = .00, partial η^2=.26), and method x test length (F(3.85, 1711.09) = 74.42, p = .00, partial η^2=.14) had a large effect size.

When the third order interactions were examined, the most important third order interaction was found to be method x average factor loading x sample size (F(7.71, 1711.09) = 346.14, p = .00, partial η^2=.61). The other third order interactions method x sample size x percentage of polytomous items (F(11.56, 1711.09) = 77.92, p = .00, partial η^2=.34), method x distribution of polytomous items x sample size (F(7.71, 1711.09) = 68.06, p = .00, partial η^2=.23), method x categories of polytomous items x sample size (F(7.71, 1711.09) = 33.01, p = .00, partial η^2=.13) had a large effect size. The other interactions were medium and small effect size, which ranged between .01-.13.

The between-subject effect was examined to investigate which simulation condition had a greater effect on the SEB values of the methods. Percentage of polytomous items had the greatest effect on SEB values (F(3, 888) = 303.42, *p* = .00, partial η^2=.51). The other simulation conditions, sample size (F(2, 888) = 144.21, *p* = .00, partial η^2=.25) and distribution of polytomous items (F(2, 888) = 104.24, *p* = .00, partial η^2=.19) had a large effect on the SEB value overall. Average factor loading (F(2, 888) = 61.69, *p* = .00, partial η^2=.12) and categories of polytomous items (F(2, 888) = 56.96, p = .00, partial η^2=.11) had a medium effect on SEB value overall. Test length (F(2, 888) = 21.09, *p* = .00, partial η^2 = .05) had a small effect on SEB value overall.

When second order interactions were examined, the most important interaction was found to be average factor loading x sample size (F(4, 888) = 65.61, *p* = .00, partial η^2 = .23) which had a large effect. The other interaction effect sizes ranged between .01-.03, and some was not statistically significant.

When post-hoc tests were examined, average factor loading categories were found to differ from each other statistically significantly: .80 had higher SEB values than .40 and .60. Similarly, .60 had higher SEB values than .40. At the same time, the condition where sample size was 1000 had statistically significantly higher SEB values than the sample size was 200. Polytomous items with 3 categories had statistically significantly smaller SEB values than those with 4 and 5 categories (*p* = .00). There was no statistically significant difference between polytomous items with 4 and 5 categories. Accordingly, the SEB values of the methods are more accurate in 4 and 5 categories polytomous items. Polytomous items which followed normal distribution had more accurate SEB values than right or left skewed ones (*p* = 00). No statistically significant difference was observed between the right or left skewed polytomous items.

The condition where test length was 20 items had more accurate SEB values than 30 and 40 item conditions (*p* = .00). No statistically significant difference was observed between the test length for 30 and 40 items. The condition where the percentage of polytomous items was 10% had more accurate SEB values than the others (10%, 20% and 40%). The increase in the percentage of polytomous items caused the SEB values to decrease. Accordingly, the decrease in the percentage of polytomous items caused more accurate SEB values.

Repeated measures ANOVA revealed that the SEB values of the methods differed from each other, and the most appropriate SEB value was in the ULSMV and MLR methods. These methods were followed by ML, WLSMV and Bayesian methods. The most effective condition of SEB values in the estimation methods was percentage of polytomous items (partial η^2 = .51). This condition was followed by sample size (partial η^2=.25), distribution of polytomous items (partial η^2=.19), average factor loading (partial η^2=.12), categories of polytomous items (partial η^2=.11) and test length (partial η^2=.01). Interaction of average factor loading x sample size (partial η^2 = .23) had the largest effect on SEB values. The effect sizes of other interactions were small (range between .01-.03).

In summary, an increase in categories of polytomous items, average factor loading, and sample size resulted in more accurate SEB values. A decrease in the test length and percentage of polytomous items resulted in more accurate SEB values. Polytomous items followed a normal distribution which makes SEB values more accurate.

3.5. Analysis of The Empirical Data Set

For sample sizes of 200, 500 and 1000 in Turkish, mathematics, science and social science tests, the convergence and inadmissible solution rates of ML/MLR and Bayesian methods were 100%

and 0%, respectively. The ULSMV method converged on all datasets but produced 8% and 6% inadmissible solutions in Turkish and mathematics datasets of 200 sample sizes, respectively. WLSMV converged in all data sets, similar to ULSMV, with 22% and 4% inadmissible solutions in Turkish and mathematics datasets with a sample sizes of 200, respectively.

When the results were examined in terms of PAE, the ML, MLR and WLS methods did not exceed 95% in any sample size. The PAE values of the Bayesian method were bigger than 95% when sample size was 1000, while it is generally below 95% when sample sizes were 200 and 500. As the average factor loading increased, the PAE values of the Bayesian method increased. The PAE values of the ULSMV and WLSMV methods were greater than 95% when the sample size was 1000. The PAE values of the WLSMV and ULSMV methods tended to increase as the average factor loading increased.

When the RB values of the estimation methods were examined, ULSMV and WLSMV methods were found to have trivial bias. The Bayesian method, on the other hand, had moderate bias only in the 200 and 500 sample sizes of the mathematics data set, and trivial bias in the other data sets.

The ML and MLR methods generally have medium bias except in the Turkish data set with a sample size of 500 and social science data set with a sample size of 200. These methods have negligible bias for these data sets.

The WLS method generally estimated the factor loadings more highly than it would if it converged. It has a large bias in 200 and 500 sample sizes. WLS has a negligible bias in the Turkish, science and social science data sets with sample sizes of 1000, however, when the PAE values of the WLS method were analyzed for these data sets, PAE values were 20%, 27% and 13%, respectively.

4. DISCUSSION and CONCLUSION

The estimation methods used for CFA in the current study were compared with mixed item response types, and thus, the performance of CFA estimation methods in mixed format tests were examined. Adding the Bayesian method as well as frequentist estimation methods allowed their performance in mixed format tests to be compared in a large number of conditions.Previous studies comparing CFA estimation methods have reported WLSMV or ULSMV methods as giving better results than estimation methods in many respects (Forero et al., 2009; Li, 2014; Rhemtulla et al., 2012; Savalei & Rhemtulla, 2013; Shi, DiStefano, McDaniel, & Jiang, 2018), however, all the items in these studies have the same number of categories.

As a result of the study, the following findings were obtained. First, the convergence rates of ML/MLR and Bayesian methods were 100% and the inadmissible solutions were 0%, similar to other studies (Forero et al., 2009; Jin, Luo, & Yang-Wallentin, 2016; Lee & Song, 2004; Li, 2016; Liang & Yang, 2014, 2016; Moshagen & Musch, 2014; Zhao, 2015). While convergence rate and inadmissible solutions of ULSMV were 100% and 0.01% respectively, WLSMV was 99.99% and 0.02%. The WLS method did not converge in small samples, as found in other studies, and the convergence rate of WLS was 49.48% and the inadmissible solution rate of WLS was 7.03% (Bandalos, 2014; Finney & DiStefano, 2013; Olsson, Foss, Troye, & Howell, 2000; Oranje, 2003).

Second, similar to other studies in the literature (Forero et al., 2009; Li, 2014; Rhemtulla et al., 2012; Savalei & Rhemtulla, 2013; Shi et al., 2018), ULSMV estimated factor loadings more accurately than other methods. Mixed item response type data thus gives similar results to non-mixed data. The WLSMV method also had similar results to ULSMV. ULSMV was more accurate in parameter estimates in this study, however, when the sample size was small (n =

200) and the average factor loading was low (.40), 8no estimation method had sufficient PAE values (PAE > 95%).

Third, when evaluated in terms of relative bias, all methods except WLS were within the acceptable range (|RB| <.10). The simulation study conducted by Shi et al. (2018) compared WLSMV, ULSMV and WLSM methods, and found that ULSMV and WLSMV methods had acceptable bias for all sample sizes (200, 500 and 1000). They also emphasized that the ULSMV method performed slightly better than the WLSMV method. Similarly, it was observed in the current study that ULSMV was less biased than other methods at a statistically significant level. The same methods were suitable in mixed item response type data. Lei (2009) found that ML and WLSMV had unbiased parameter estimates. The estimation methods gave similar results in mixed item response type data to five point categorical data. Liang and Yang (2014) stated that the WLSMV method is slightly better than the Bayes method in terms of bias. Since non-informative priors were used in the current study, the Bayesian method may have had a larger bias than other methods, however, the RB value of the Bayesian method was also within the acceptable range (|RB| <.10).

Forth, the standard error bias (SEB) values of all methods, except WLS, were negligible with increasing sample size. The SEB values of all methods are acceptable, except those for WLS, however, the SEB values differed statistically significantly according to the methods. The ULSMV and MLR methods had the least SEB value. Repeated measures ANOVA demonstrated that the SEB values of the methods differed from each other, and that the ULSMV and MLR methods had the most appropriate SEB value for all simulation conditions. Generally, the increase of categories of polytomous items, factor loading, and sample size make the SEB value more accurate, and the decrease in the test length, the percentage of polytomous items and polytomous items follow normal distribution and make the SEB values more accurate. Jin et al. (2016) also noted that the SEB values of WLSMV, ULS and ML methods were acceptable. Mixed item response type data does not cause a big change in the SEB values of the methods.

Similar results to those of the simulation study were obtained in the analyses performed with real data sets. ML/MLR and Bayesian methods converged in all datasets and had no inadmissible solution. ULSMV and WLSMV converged in all datasets but had a small number of inadmissible solutions. All methods, except WLS, had acceptable RB values.

In conclusion, the ULSMV estimation method is preferable when performing CFA with mixed item response type data, so that parameter estimates can be more accurate. Although the results of the methods are within the acceptable range in terms of RB and SEB values, when evaluated in terms of PAE, ULSMV is slightly better than WLSMV in parameter estimates. However, it should be remembered that the method's PAE values were not in the acceptable range for small sample sizes and low average factor loading. No estimation method is suitable for every condition in mixed item response type data, and the estimation method should be selected considering the sample size and the average factor loading. In future studies, researchers could perform simulation studies manipulating the number of factors, and correlations between factors, using informative priors for the Bayesian method. This study is limited to MEAS data sets collected in 2016. This study is also limited to the 491 simulation conditions at the time. In the current study, mixed item response type data was created to be binary and three categories, binary and four categories or binary and five categories polytomous data independently. This could be manipulated in future studies as binary and three, four and five categories, simultaneously.

Acknowledgments

This paper was produced from the first author's doctoral dissertation.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship contribution statement

Abdullah Faruk KILIÇ: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Nuri DOĞAN:** Investigation, Methodology, Supervision, and Validation.

ORCID

Abdullah Faruk KILIÇ https://orcid.org/0000-0003-3129-1763
Nuri DOĞAN https://orcid.org/0000-0001-6274-2016

5. REFERENCES

AERA, APA, NCME, American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement In Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *24*(2), 222-228. https://doi.org/10.2307/3151512

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 102-116. https://doi.org/10.1080/10705511.2014.859510

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 625-666). Information Age.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118619179

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Depaoli, S., & Scott, S. (2015). Frequentist and bayesian estimation of CFA measurement models with mixed item response types: A monte carlo investigation. *Structural Equation Modeling: A Multidisciplinary Journal*, (September), 1-16. https://doi.org/10.1080/10705511.2015.1044653 (Retraction published 2015, Structural Equation Modeling: A Multidisciplinary Journal, 318)

DiStefano, C. (2002). The impact of categorization with confirmatory cactor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(3), 327-346. https://doi.org/10.1207/S15328007SEM0903_2

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241. https://doi.org/10.1177/073428290502300303

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309–326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

Ferguson, E., & Rigdon, E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, *28*(4), 491–497. https://doi.org/10.2307/3172790

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Information Age.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 625-641. https://doi.org/10.1080/10705510903203573

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 108–120. https://doi.org/10.1080/10705519709540064

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*(4), 427–438. https://doi.org/10.1177/001316444600600401

Hallquist, M., & Wiley, J. (2017). *MplusAutomation: Automating Mplus model estimation and interpretation*. Retrieved from https://cran.r-project.org/package=MplusAutomation

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A comparison of ML, WLSMV, and bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research*, *51*(5), 661-680. https://doi.org/10.1080/00273171.2016.1208074

Jin, S., Luo, H., & Yang-Wallentin, F. (2016). A simulation study of polychoric instrumental variable estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 680-694. https://doi.org/10.1080/10705511.2016.1189334

Lee, T. K., Wickrama, K., & O'Neal, C. W. (2018). Application of latent growth curve analysis with categorical responses in social behavioral research. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 294-306. https://doi.org/10.1080/10705511.2017.1375858

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*(4), 653–686. https://doi.org/10.1207/s15327906mbr3904_4

Lei, P. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, *43*(3), 495–507. https://doi.org/10.1007/s11135-007-9133-z

Li, C.-H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables* [Unpublished Doctoral dissertation]. Michigan State University.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of*

Quantitative Research in Education, *2*(1), 17-38. https://doi.org/10.1504/IJQRE.2014.060972

Liang, X., & Yang, Y. (2016). Confirmatory factor analysis under violations of distributional and structural assumptions: A comparison of robust maximum likelihood and bayesian estimation methods. *Journal of Psychological Science*, *39*(5), 1256–1267. https://doi.org/10.1504/IJQRE.2013.055642

Lorenzo-Seva, U., & Ferrando, P. J. (2020). *Factor* (Version 10.10.03) [Computer software]. Universitat Rovira i Virgili.

MoNE. (2017). *Monitoring and evaluation of academic skills report for eight graders*. MONE. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf

Morata-Ramirez, M. de los A., & Holgado-Tello, F. P. (2013). Construct validity of likert scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *International Journal of Social Science Studies*, *1*(1), 54-61. https://doi.org/10.11114/ijsss.v1i1.27

Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, *10*(2), 60-70. https://doi.org/10.1027/1614-2241/a000068

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, B. O., & Kaplan, D. (1992). 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*, *45*(1), 19–30. https://doi.org/10.1111/j.2044-8317.1992.tb00975.x

Muthén, L. K., & Muthén, B. O. (2012). *Mplus statistical modeling software: Release 7.0* [Computer software]. Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd. ed.). McGraw-Hill.

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: A Multidisciplinary Journal*, *7*(4), 557–595. https://doi.org/10.1207/S15328007SEM0704_3

Oranje, A. (2003, April 21-25). *Comparison of estimation methods in factor analysis with categorized variables: Applications to NEAP data* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education. Chicago, IL, USA.

Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS®*. SAS Intitute Inc.

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, *46*(2), 273–286. https://doi.org/10.1111/j.2044-8317.1993.tb01016.x

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/.

Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research*. https://cran.r-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201–223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2018). Examining chi-square test statistics under conditions of large model size and ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 924-945. https://doi.org/10.1080/10705511.2018.1449653

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, *56*(2), 197–208. https://doi.org/10.1177/0013164496056002001

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *National Institutes of Health*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(3), 392–423. https://doi.org/10.1080/10705511.2010.489003

Zhao, Y. (2015). The performance of model fit measures by robust weighted least squares estimators in confirmatory factor analysis [Doctoral dissertation, The Pennsylvania State University]. https://etda.libraries.psu.edu/catalog/24901