The influence of model size on the estimation accuracy of estimation methods in structural equation models with ordinal variables

Andreas Falke
Regensburg University

Cite as:
Falke Andreas (2019), The influence of model size on the estimation accuracy of estimation methods in structural equation models with ordinal variables. *Proceedings of the European Marketing Academy*, 48th, (8102)

Paper presented at the 48th Annual EMAC Conference, Hamburg, May 24-27, 2019.

The influence of model size on the estimation accuracy of estimation methods in structural equation models with ordinal variables

Structural equation modeling has become a popular tool in marketing but a problem with the its application is that most researchers use ordinal answer scales in their surveys, whereas most of the popular estimation methods assume continuous variables. Estimation methods that can deal with ordinal scales have been published; however, the impact of model size on estimation accuracy of these methods has not been investigated. This study uses a Monte Carlo simulation to test, how well five different estimation methods (three that assume continuous variables, two that can deal with ordinal variables) perform under several model size constellations. Apart from estimation method and model size, sample size and two factors on construct validity are also considered. Results show that diagonally weighted least squares with a polychoric correlation matrix is among the best estimation methods most of the time, but, in several constellations, other estimation methods often perform equally well.

keywords: structural equation modeling, estimation method, ordinal variables
track: Methods, Modelling and Marketing Analytics

1 Introduction

Since its introduction over 35 years ago, structural equation modeling has become one of the most popular tools in marketing and social sciences (see, e.g., Hwang, Malhotra, Kim, Tomiuk, and Hong, 2010). But there are several problems associated with structural equation modeling. A serious problem occurs during the data collection process. Often, subjects are given ordinal answer scales, e.g., Likert-type scales, but during the estimation process, these items are treated as continuous variables because most of the widely used estimation methods such as maximum likelihood (ML) assume continuous and multivariate normally distributed variables (Bollen, 1989). This mistreatment undermines the accuracy of model estimates and can lead to wrong or at least misleading conclusions (Kaplan, 2009). Estimation methods for ordinal scales are available, but studies comparing these methods provide mixed findings. The main research goal of this paper consists in clarifying how different estimation methods influence estimation accuracy for structural equation models of different model sizes. Three different estimation methods that are commonly used (conventional ML, ULS, and GLS) as well as two methods that are developed for ordinal variables (ULS-cat and DWLS-cat) are investigated.

The paper differs from the extant literature on structural equation modeling simulation with ordinal variables in three important aspects. First, this study simulates the values of each artificial respondent individually. To the knowledge of the author, no other simulation study (e.g., Marsh, Hau, Balla, and Grayson, 2004) tried to accomplish this. Second, model size is split into three different experimental factors: the number of exogenous constructs, number of endogenous constructs, and the number of items per construct. Prior studies (e.g., Li 2016) often use a fixed model or only include a single factor for model size. An exception is the recent study of Shi, DiStefano, McDaniel, and Jiang (2018) who differentiate between the number of items and the number of constructs. As a third distinguishing aspect, this study also considers construct reliability as an influence on estimation accuracy. Most of the time, studies use factor loadings as experimental factor (see, e.g., Reinartz, Haenlein, and Henseler, 2009) and do not analyze if the resulting constructs vary in their reliability.

2 Structural equation modeling

Consistent with literature (Bollen, 1989), the usual notation for the structural model is applied:

$$\eta = B\eta + \Gamma\xi + \zeta \tag{1}$$

where η and ξ are the endogenous and exogenous constructs, respectively, and ζ is a vector of random errors. The reflective measurement models of the SEM are defined as:

$$x = \Lambda_x\xi + \delta \qquad y = \Lambda_y\eta + \varepsilon. \tag{2}$$

x and y stand for the p- and q-dimensional vectors of manifest items, Λ_x and Λ_y are item loading matrices and δ and ε denote measurement error vectors.

3 Experimental Design of the Monte Carlo study

3.1 Experimental Factors

Five different methods to estimate the structural equation models are tested:

1. Maximum likelihood (ML) estimation
2. Unweighted least squares (ULS) estimation
3. Generalized least squares (GLS) estimation
4. ULS estimation with a polychoric correlation matrix (ULS-cat)
5. Diagonally weighted least squares estimation with a polychoric correlation matrix (DWLS-cat)

Studies have shown that ULS-cat and DWLS-cat often outperform other estimation methods (e.g. Li, 2016), but the results are not unanimous in whether ULS-cat or DWLS-cat is superior (Bandalos, 2014, Shi et al., 2018). Furthermore, the studies state that ML, ULS, and GLS also are not without merits. Apart from the estimation method, six additional experimental factors are used in this study (see Table 1). Model size is a source of model fit variation (Kenny & McCoach, 2003). More items lead to a higher-dimensional covariance matrix and can be used to gain more degrees of freedom. Furthermore, model size effects can stem from the increase in the model's latent constructs (e.g., Breivik & Olsson, 2001) and the increase in items per latent construct (e.g., Marsh et al., 1998). As model fit variation can stem from any of these sources, model size is divided into three factors: the number of exogenous and endogenous constructs and the average number of items per construct.

Table 1: Experimental Factors in the Simulation

Experimental Factor	Variable Name	Factor Levels
Estimation method	estiMet	ML, ULS, GLS, ULS-cat, DWLS-cat
Model size		
Number of exogenous constructs	numExo	1, 2, 3
Number of endogenous constructs	numEndo	1, 2, 3
Average number of items per construct	aveItems	3, 4
Sample size	sample	100, 300, 500
Construct reliability		
Cronbach's alpha level per construct	alphaMean	low, high
Variance of Cronbach's alpha level	alphaVar	none, low, high

Sample size is is commonly used experimental factor in simulation studies (see, e.g., Li, 2016), so it is also considered here.

Most of the time, homogeneous factor loadings are common in simulation studies (e.g., Flora & Curran, 2004), but they are not common outside of simulations. Instead of fixating factor loadings, the focus is set on Cronbach's alpha levels. The reason for this choice is that researchers use alpha level of constructs as criterion for the fit of the measurement model. Therefore, it is of interest to see if the alpha level also influence the estimation method.

3.2 Estimation Accuracy

In this study, estimation accuracy of structural equation models refers to the discrepancy between estimated and true model components and is measured by computing mean absolute relative errors (MARE). Using only one MARE as measure would be too broad, so not only one, but three different MARE values are computed for each model: One value that considers all path coefficients and factor loadings ($MARE_{total}$), one value that considers only the path coefficients ($MARE_{path}$), and one value that considers only the factor loadings ($MARE_{factor}$).

3.3 Generating Simulated Data

In the Monte Carlo study 324 different constellations are considered and, in total, 153,000 models are estimated. The overall simulation consists of the following steps: First, the true values for each exogenous and endogenous construct for each respondent *i* from the sample are calculated. Then, based on these constructs, the item values for each respondent *i* are calculated. These items are rescaled and discretized to fit on a 7-point Likert-scale. Finally, the components are estimated using all five estimation types and the three MARE values between the true and estimated components are computed.

4 Major Findings

Two ANOVAs are performed: one with the three different MARE variants and one with the logarithm of each MARE variant as the dependent variable. The R^2 values were always higher in the latter case, so ln(MARE) is chosen as evaluation criteria. A measure of prediction accuracy like ln(MARE) as dependent variable in an ANOVA results in negative coefficient showing decreasing effects on estimation error. On the other hand, a positive coefficient points toward an increasing estimation error. Furthermore, all factor coefficients have to be interpreted relative to the respective reference levels, which are the first factor levels written in Table 1. Table 2 depicts the main effects and interactions of the ANOVA. As a reading example: The positive coefficients of estiMet ULS in Table 2 means that models estimated with ULS have on average a higher total estimation error, higher errors in the path coefficients, and higher errors in the factor loadings compared to models estimated with ML.

4.1 Main effects

First, the main effects concerning $ln(MARE_{total})$ are described. The main effects of estimation methods display only minor differences on estimation error: The use of ULS-estimation

Table 2: Analysis of variance: coefficients of $ln(MARE_{total})$ / $ln(MARE_{path})$ / $ln(MARE_{factor})$

	Predictor	Coefficients	Predictor	Coefficients	Predictor	Coefficients
Main effects	(Intercept)	**-1.83 / -1.65 / -1.96**	numExo 2	**0.12 / 0.34 / -0.02**	sample 400	**-0.14 / -0.16 / -0.15**
	estiMet ULS	**0.06 / 0.06 / 0.03**	numExo 3	**0.20 / 0.53 / -0.05**	sample 500	**-0.24 / -0.26 / -0.27**
	estiMet GLS	0.01 / 0.01 / 0.01	numEndo 2	**0.49 / 1.09 / 0.04**	alphaMean high	**-0.51 / -0.30 / -0.72**
	estiMet ULS-cat	**0.02 / 0.10 / -0.05**	numEndo 3	**0.67 / 1.31 / 0.08**	alphaVar low	**-0.03 / -0.05 / -0.04**
	estiMet DWLS-cat	0.00 / **0.10 / -0.07**	aveItems 4	**-0.02 / 0.05 / 0.06**	alphaVar none	**-0.03 / -0.03 / -0.05**
estiMet interactions		ULS	GLS	ULS-cat	DWLS-cat	
	numExo 2	0.00 / **-0.04 / 0.05**	**0.02 / 0.02 / 0.02**	-0.01 / **-0.07 / 0.06**	**-0.02 / -0.07 / 0.03**	
	numExo 3	0.00 / **-0.08 / 0.08**	**0.04 / 0.03 / 0.04**	-0.01 / **-0.10 / 0.10**	**-0.03 / -0.10 / 0.06**	
	numEndo 2	**-0.07 / -0.10** / -0.01	**0.01** / 0.00 / **0.03**	**-0.07 / -0.17** / -0.01	**-0.06 / -0.16** / 0.00	
	numEndo 3	**-0.09 / -0.11 / -0.02**	**0.05 / 0.04 / 0.07**	**-0.08 / -0.17 / -0.02**	**-0.06 / -0.16** / 0.00	
	aveItems 4	0.00 / -0.01 / -0.02	**0.01** / 0.00 / **0.02**	**-0.01** / -0.01 / **-0.03**	-0.01 / -0.01 / -0.01	
	sample 400	-0.00 / -0.00 / -0.00	**-0.01** / -0.01 / **-0.02**	0.00 / 0.01 / 0.00	0.00 / -0.01 / 0.00	
	sample 500	-0.01 / -0.00 / 0.00	**-0.02** / -0.01 / **-0.03**	0.00 / 0.02 / 0.00	0.01 / 0.02 / 0.00	
	alphaMean high	-0.00 / **-0.02 / -0.06**	-0.01 / 0.01 / 0.00	**-0.03** / 0.01 / **-0.05**	**-0.06** / 0.01 / **-0.13**	
	alphaVar low	0.00 / -0.00 / 0.01	-0.00 / -0.00 / 0.00	**-0.01 / -0.04** / 0.00	**-0.01 / -0.04** / 0.00	
	alphaVar none	-0.00 / -0.01 / 0.00	-0.00 / -0.00 / -0.00	**-0.02 / -0.05** / -0.00	**-0.02 / -0.05** / 0.00	
Model size interactions		numExo 2	numExo 3	numEndo 2	numEndo 3	aveItems 4
	numEndo 2	**-0.26 / -0.64** / 0.01	**-0.43 / -0.99** / -0.01			
	numEndo 3	**-0.28 / -0.59 / -0.02**	**-0.47 / -0.98** / 0.01			
	aveItems 4	**0.02 / -0.02 / 0.01**	**0.04** / -0.01 / **0.02**	**-0.07** / -0.01 / -0.01	**-0.10** / -0.01 / **-0.01**	
	sample 400	-0.01 / **0.02** / -0.00	**-0.03 / -0.03** / 0.00	**0.04 / 0.10 / -0.01**	**0.05 / 0.10 / -0.02**	**-0.02 / -0.02 / -0.02**
	sample 500	**-0.03** / -0.01 / 0.01	**-0.05 / -0.05** / 0.00	**0.09 / 0.19** / -0.00	**0.11 / 0.20 / -0.01**	**-0.02** / -0.01 / **-0.02**
	alphaMean high	**-0.03 / -0.03 / 0.02**	**-0.04 / -0.07 / 0.06**	**0.20 / 0.25** / 0.00	**0.26 / 0.29 / -0.03**	**-0.07 / -0.03 / -0.01**
	alphaVar low	-0.00 / -0.01 / 0.00	**-0.01 / -0.02** / -0.00	**0.01 / 0.04** / 0.00	**0.02 / 0.06** / 0.01	**-0.01** / -0.01 / **-0.02**
	alphaVar none	-0.00 / **-0.02 / 0.01**	**-0.01 / -0.02** / 0.00	**0.01 / 0.05** / 0.00	**0.02 / 0.05** / 0.01	-0.00 / **-0.02** / 0.01
Other interactions	sample 400 × alphaMean high			**0.04 / 0.03 / 0.02**		
	sample 500 × alphaMean high			**0.04 / 0.02 / 0.02**		
	sample 400 × alphaVar low			**0.02 / 0.02 / 0.02**		
	sample 500 × alphaVar low			**0.04 / 0.02 / 0.02**		
	sample 400 × alphaVar none			**0.01** / -0.00 / **0.01**		
	sample 500 × alphaVar none			**0.01** / 0.00 / **0.02**		
	alphaMean high × alphaVar low			**0.02** / 0.01 / **0.03**		
	alphaMean high × alphaVar none			**0.02** / 0.01 / **0.03**		

bold indicates significance at $\alpha = 0.05$

$R^2 = 0.62 / 0.46 / 0.57$

Reading example: The main effect of ULS is 0.06 if $ln(MARE_{total})$ is dependent variable, 0.06 for $ln(MARE_{path})$, and 0.03 for $ln(MARE_{factor})$

and surprisingly ULS-cat are slightly worse than any other estimation method, which do not differ significantly from each other. Using more constructs will lead to higher estimation errors. According to the standardized beta coefficients, this effect is much more prominent when the number of endogenous constructs is increasing than in case of an increase of the number of exogenous constructs. As expected, higher sample sizes lead to smaller estimation errors and so does a higher mean level of Cronbach's alpha. Interestingly, lower or no variances also induce a smaller estimation error.

When analyzing the main effects of $ln(MARE_{path})$, ULS-cat und DWLS-cat have the highest main effect, closely followed by ULS. Using these methods leads to a higher estimation error while GLS and ML do not differ significantly. All coefficients concerning model size are significantly positive, which means larger models tend to have larger estimation errors, regardless if there are more constructs or more items per construct. Again, the effect is more pronounced in case of more endogenous constructs. Furthermore, as expected, larger sample sizes and higher alphaMean levels with low variance lead to smaller estimation errors.

With the exception of estiMet, the main effects of $ln(MARE_{factor})$ are in line with the other ANOVAS. Here, ULS-cat and DWLS-cat have the smallest estimation error, ULS the highest, and ML and GLS are in the middle. The main effects of numEndo and aveItems have the same sign as in the other analyses, but numExo has negative coefficient, i.e., using more exogenous constructs reduces the estimation error of factor loadings. Using a higher sample size and a higher Cronbach's alpha mean with low variance also yields smaller estimation errors.

4.2 estiMet interactions

Many interactions of the estiMet levels with other factors are significant, which shows that estimation methods react differently to these factors. ULS is the estimation method that is influenced the least by other experimental factors. Here, higher numbers of exogenous items lead to lower estimation errors among the path coefficients, but higher errors among the factor loadings. In contrast, more endogenous constructs always yield lower estimation errors in all parts of the model. If GLS is used, every kind of estimation error gets bigger with larger models. For ULS-cat and DWLS-cat, the interactions with other experimental factors are very similar. Larger models lead to a lower estimation error among path coefficients and higher estimation errors among factor loadings. As shown by $ln(MARE_{total})$, the overall effect of larger models is a reduction of the estimation error. Both ULS-cat and DWLS-cat also interact with the construct reliability factors. Interestingly, while a higher mean of Cronbach's alpha leads to lower estimation errors among the factor loadings, lower or no variances in the alphas lead to lower estimation errors among path coefficients.

4.3 Model size interactions

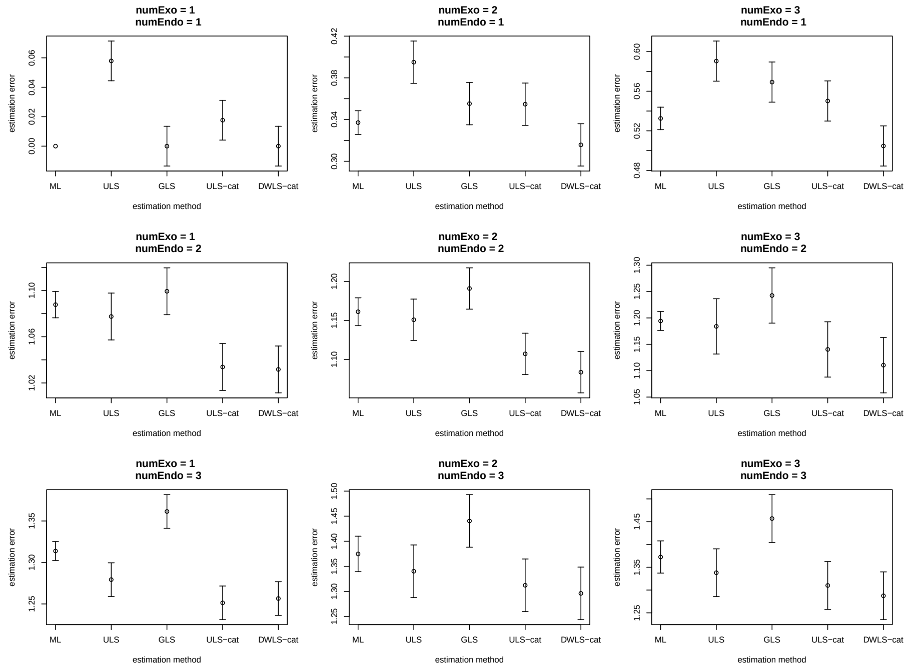
The next part of Table 2 depicts the interactions of the model size factors with the other factors. The interactions between the number of different construct types are significant and have the opposite sign as both of their respective main effects. As a consequence, an increase of model size has subadditive effects, i.e., the estimation error will rise but less than linear. Furthermore, the interactions of numExo are negative for errors among path coefficients and positive for errors among factor loadings, while the interactions of numEndo carry the opposite sign and have a larger absolute value. This shows that the estimation errors react very differently if the number of either type of constructs is altered.

4.4 Construct reliability

The general result is that a high Cronbach's alpha level supports good estimation accuracy. That in itself is not surprising. Interestingly, having a low or no variance in alpha levels also helps in reducing the estimation error, albeit this effect is not as strong as having a high mean. For path coefficients, the reduction of estimation errors from low or no alpha level variance is larger if ULS-cat or DWLS-cat is used. Furthermore, alphaMean often interacts with model size factors. In models with many exogenous constructs, having a high mean alpha level and/or a low alpha level variance can improve the estimation errors on path coefficients. In models with many endogenous constructs, on the other hand, the alpha level only has a small or no influence on estimation accuracy. When estimation errors of factor loadings are important, for instance in confirmatory factor analyses, then having a high alpha level is essential, as can be seen by the coefficient. Most of the estimation methods (except ML and GLS) further strengthen this influence. The impact of a high alpha level is reduced in models with many exogenous factors and enhanced in models with many endogenous factors.

4.5 Total effects of the estimation methods

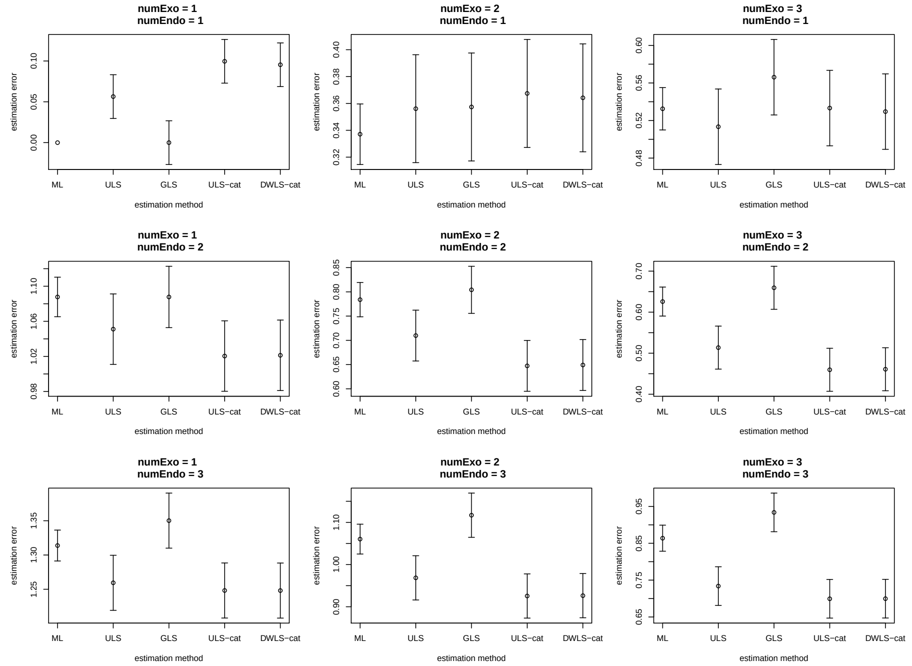
Figure 1: 95% confidence intervals of total effects concerning $ln(MARE_{total})$



As can be seen in Table 2, the estimation methods often interact with the different model size factors. Furthermore, the signs and absolute values of these interactions vary to a large extent. To get a whole picture of these scenarios is difficult, so the total effects, i.e., the sum of the main effects and significant pairwise interactions, of estiMet, numExo, and numEndo, are investigated. The factor aveItems is not varied because its interactions are often small or not significant. All other factors are kept at their respective reference categories. Figures 1, 2, and 3 show the 95% confidence intervals of the total effects under several different model size factor combinations. The lower the estimation errors of a total effect, the more appropriate is the corresponding estimation method in the constellation.

Figure 1 shows the 95% confidence intervals concerning $ln(MARE_{total})$. In case of only one exogenous and one endogenous construct, ML, GLS as well as DWLS-cat are the best possible estimation methods, because their confidence intervals indicate that the estimation errors are the smallest ones. If the number of endogenous constructs are kept constant at 1 and the number of exogenous constructs increase (first row), ML and GLS become worse. As a conse-

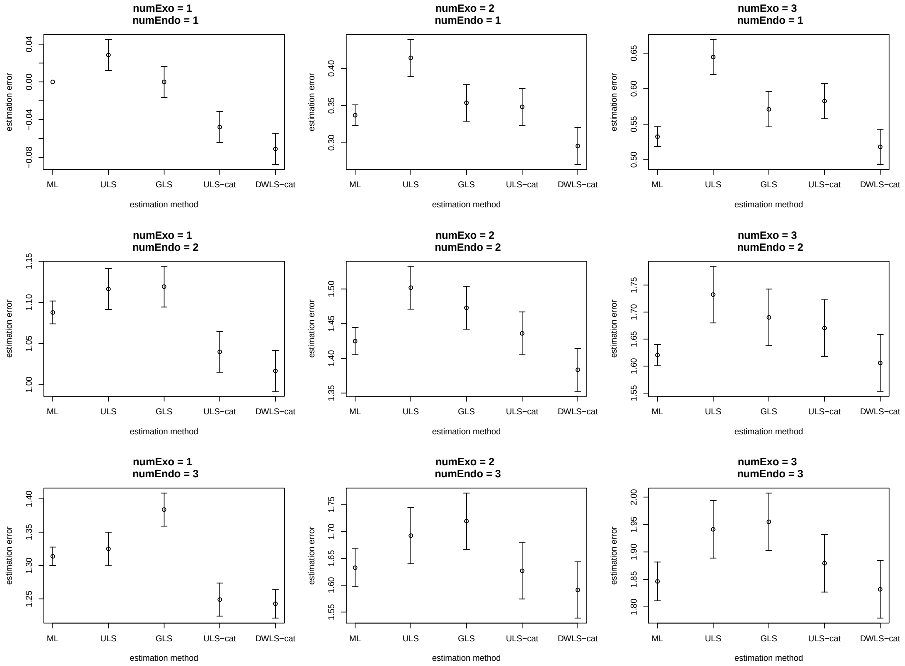
Figure 2: 95% confidence intervals of total effects concerning $ln(MARE_{path})$



quence, in these constellations, DWLS-cat should be used. If the model consists of more than 1 endogenous construct, but less than 3 exogenous constructs (the four constellations in second and third row and first and second column), than ULS-cat and DWLS-cat are clearly the preferred estimation methods. If the model consists of 3 exogenous and more than 1 endogenous construct (third column), it is not easy to give a recommendation. DWLS-cat is clearly better in both constellation than GLS, because both confidence intervals are disjoint. This is also the case for ULS-cat and GLS in case of 3 exogenous and 3 endogenous constructs. As a consequence, the application of DWLS in both cases and ULS-cat in case of 3 endogenous constructs would be recommended.

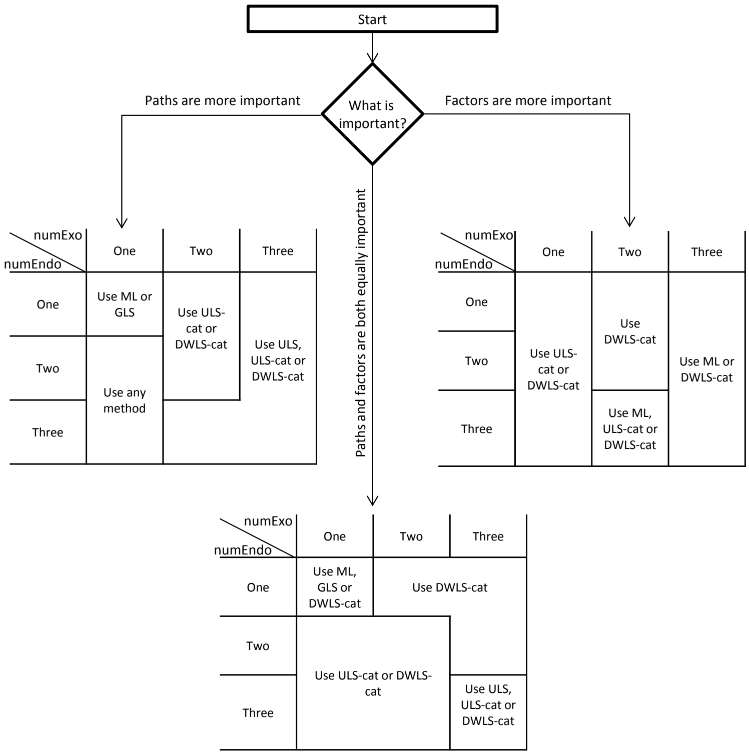
In Figure 2, the estimation error of path coefficients is the dependent variable. One clearly distinguished case is the scenario of only 1 exogenous and 1 endogenous construct. Here, ML and GLS are the best possible estimation methods, because their confidence intervals are disjoint from those of the other estimation methods while retaining the lowest estimation errors. On the other hand, when only the number of exogenous constructs increases (first row), the estimation methods are not distinguishable, because all confidence intervals overlap. The scenario with 1 exogenous and 2 endogenous construct is also ambiguous; it can only be said, that ML is worse than ULS-cat or DWLS-cat. If the number of both construct types increases, the differences between the estimation methods become more pronounced. ULS-cat and DWLS-cat both become the best estimation methods, while ML and GLS become the worst to apply. The accuracy of ULS is highly dependent on the combination. If numEndo = 2, then the estimation errors of ULS is between ML/GLS and ULS-cat/DWLS-cat, but if numEndo = 3, then ULS is not distinguishable from ULS-cat or DWLS-cat.

Figure 3: 95% confidence intervals of total effects concerning $ln(MARE_{factor})$



In Figure 3, $ln(MARE_{factor})$ is the dependent variable. If there is only one exogenous construct (first column), ULS-cat and DWLS-cat are clearly superior to the other three estimation methods. ML, ULS and GLS do not differ much (numEndo = 1 or 2), or ML and ULS are better than GLS (numEndo = 3). If the number of exogenous constructs is 2 (middle column), DWLS-cat is the best estimation method. The other estimation methods perform similar to each other, as their confidence intervals overlap most of the time. Only ML performs clearly better than ULS (numEndo = 1 or 2) or GLS (numEndo = 3). Finally, in the last column, ML und DWLS-cat have the lowest estimation errors. Overall, ULS is almost always among the estimation method with the highest estimation error, while DWLS-cat (and ML if numExo = 3) is among the best.

Figure 4: Guideline for choosing appropriate the estimation method


Start
What is important?
Paths are more important
Factors are more important
Paths and factors are both equally important

numEndo \ numExo	One	Two	Three
One	Use ML or GLS	Use ULS-cat or DWLS-cat	Use ULS, ULS-cat or DWLS-cat
Two	Use any method		
Three			

numEndo \ numExo	One	Two	Three
One	Use ULS-cat or DWLS-cat	Use DWLS-cat	Use ML or DWLS-cat
Two			
Three		Use ML, ULS-cat or DWLS-cat	

numEndo \ numExo	One	Two	Three
One	Use ML, GLS or DWLS-cat	Use DWLS-cat	
Two	Use ULS-cat or DWLS-cat		
Three			Use ULS, ULS-cat or DWLS-cat

5 Conclusions

This study evaluates the effects of five different estimation methods on estimation accuracy in structural equation models and the interaction with model size. ML, ULS, GLS, ULS-cat, and DWLS-cat are investigated and used as one experimental factor while model size is split into three factors. Furthermore, several other important aspects of structural equation modeling are also captured in experimental factors: sample size, and the construct reliability. To evaluate the estimation accuracy, three kinds of MARE are used. The results show that with different model sizes different estimation methods are optimal. An analysis of the total effects delivers different appropriate estimation methods for different model sizes. Figure 4 summarizes these findings and should guide researchers to the best estimation method of their model. Most of the time, DWLS-cat is among the best estimation methods. But for the majority of models, other estimation methods perform as well as DWLS-cat. For larger models models, ULS (ML) can be a viable alternative if path coefficients (factor loadings) are of the essence for the researcher.

Key References

Bandalos, D. (2014). Relative performance of categorial diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 102-116.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley-Interscience.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.

Kenny, D.A., & McCoach, D.B. (2003). Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(3), 333-351.

Li, C.H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, 21(3), 369-387.

Marsh, H.W., Hau, K.T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2), 181-220.

Reinartz, W., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, 26(4), 332-344.

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*, Article in Press.