Scale Validation Conducting Confirmatory Factor Analysis: A Monte Carlo Simulation Study With LISREL

Daniel Ondé* and Jesús M. Alvarado

Psychobiology and Behavioral Sciences Methods, Complutense University of Madrid, Madrid, Spain

INTRODUCTION

When psychologists are going to test their theoretical models (at the time of planning the research study), several questions may arise regarding the quality and potential accuracy of the estimation of Confirmatory Factor Analysis (CFA) models under certain applied conditions. For example, one question is the minimum sample size (N) and/or the number of indicators per factor (p/k) that is needed to estimate the CFA models properly. Many of these questions can be answered through simulation studies, because the magnitudes of the population factor loadings (λ_{ik}) are known in advance. Monte Carlo simulation uses random sampling and statistical modeling to estimate mathematical functions, and is a key tool for studying analytically intractable problems (Harrison, 2010). It is quite frequent to find in the literature simulation studies that use CFA to fit measurement models. However, there is a lack of technical information in the published research to replicate this type of studies, probably due to length limitations. Furthermore, researchers and scholars must often go to numerous and technically complex sources of information to understand the laborious simulation and estimation process.

In this paper we present complete technical information to conduct Monte Carlo simulation CFA-studies with PRELIS and LISREL programs. The LISREL program, apart from being one of the most used (and validated) software programs, is historically linked to CFA (Brown, 2015). Although it is a commercial program, there is a free student version that allows performing all the simulation tools and the analysis techniques shown in this work. Through a simulation study, we have evaluated the necessary conditions to test the CFA models fit, so we have chosen population structures formed by a single common factor to design the simulation study and low-moderate population factor loadings. It should be noted that these types of studies generate a considerable volume of information, even when using a simple unifactorial model. This approach was adopted for this study for ease of understanding the principles. Included, however, are the PRELIS and LISREL syntax examples for multidimensional factor structures. In the Supplementary Material (https://figshare.com/s/18eb0e998150d39bc952) we have attached both the simulated data and the CFA results (parameter estimates, standard errors and goodness-of-fit measures).

MONTE CARLO SIMULATION STUDY

Simulated Conditions

The simulated experimental conditions reflect low-medium sample sizes ($N = 200, 300, 400,$ and 500), low-moderate model sizes (indicators per factor: $p/k = 4, 5, 6, 7,$ and 15), and low-moderate population factor loadings ($\lambda_{ik} = 0.2, 0.3,$ and 0.4). The unifactorial structures simulated

are equal-λ_{ik} condition (i.e., all p/k indicators of each population structure have the same λ_{ik} magnitude). The selected sample-model size is frequent in applied research. For example, after reviewed 1,409 published CFA models, Jackson et al. (2009) found a median N of 389. On the other hand, a common recommendation in some applied context is not to interpret estimated factor loadings < 0.30 (Brown, 2015).

Data Generation

The data generation process was conducted using PRELIS 2 program (Jöreskog and Sörbom, 1996b), according to the common factor model showed in Equation (1):

$$\Sigma = \Lambda \Phi \Lambda^{'} + \Theta \qquad (1)$$

were Σ is the population correlation matrix, Λ is the population factor loading matrix, Φ is the population factor correlation matrix, and Θ is the unique variances matrix. All indicators were simulated as continuous variables, computed by the expression shown in Equation (2):

$$\mathbf{x} = \Lambda \boldsymbol{\xi} + \boldsymbol{\delta} \qquad (2)$$

Equation (2) summarized a set of equations (one for each indicator: $X_i = \lambda_{ik} \boldsymbol{\xi}_k + \delta_i$) that express the relationship among the observed variables (X_i), the common factor ($\boldsymbol{\xi}_k$) and the unique variances (δ_i).

Step 1 of the data generation process has consisted, following Equation 2, in the simulation of the common factor ($\boldsymbol{\xi}_k$) as a random variable (in this case, $\boldsymbol{\xi}_k = 1$, see Jöreskog and Sörbom, 1996b; Brown, 2015), and the computation of the p/k indicators. Following the recommendation of Kelley and Pornprasertmanit (2016), we have simulated 1,000 normal distributed data matrix $\mathbf{X}$ (order N x p/k indicators), with the same integer starting value for the random number generator in all the simulated conditions (see Example 1 from Presentation 1). To extend the experimental simulation conditions, Example 1 shows how to recode continuous variables into discrete variables using thresholds, and how to simulate non-normal distributed indicators.

For each data matrix $\mathbf{X}$ generated in step 1, step 2 has consisted in the generation of files that contain variance-covariance matrices ($\mathbf{S}$) and files that contain correlation matrices ($\mathbf{R}$). That is, both types of $\mathbf{R}$ and $\mathbf{S}$ files come from the same raw data (see Example 2 from Presentation 1).

Estimation of CFA Model Parameters

The objective of CFA is obtaining estimates for each parameter of the measurement model and computing a predicted variance-covariance or correlation matrix (Σ^*) that best reproduces the observed or input matrix ($\mathbf{S}$ or $\mathbf{R}$). In practice, perfect fit is not possible due to the measurement error, so the starting point of analysis implies that $\Sigma^* \neq \mathbf{S}$ or $\mathbf{R}$ (Bollen, 1989). The residuals (d_i) are the differences between each pair of predicted-observed covariances or correlations, so vector $\mathbf{d} = \{d_1, d_2,..., d_i\}$ indicates the degree of discrepancy between both matrices. The estimation process consists in an iterative numerical technique to minimize the discrepancy between Σ^* and $\mathbf{S}$ (or $\mathbf{R}$) matrices (Bollen, 1989; Brown, 2015). This numerical technique is a lineal fitting function [also called discrepancy function $F(\mathbf{S}; \Sigma^*)$ or $F(\mathbf{R}; \Sigma^*)$], which takes different forms depending on the estimation method used.

Maximum Likelihood (ML) fitting function is widely used in applied CFA research (Brown, 2015). ML minimization function can be expressed as $F_{ML} = \mathbf{d'Wd}$ (where $\mathbf{W}$ is a weighted matrix), where F is calculated from each computed vector $\mathbf{d}$ at each step of the iterative process. Unweighted Least Square (ULS) is an easy to understand fitting function, $F_{ULS} = \mathbf{d'd}$, where F is obtained by calculating the sum of squares of the residuals d_1, $d_2,..., di$ ($\mathbf{W}$ can be replaced by an identity matrix ($\mathbf{I}$), simplifying the ML discrepancy function). F_{ULS} estimates are consistent estimates with sufficient sample size, although less efficient than F_{ML} estimates (Bollen, 1989).

A cautionary note should be made regarding the use of $\mathbf{S}$ and $\mathbf{R}$ matrices. The general rule is that $\mathbf{S}$ should be analyzed, although a common practice is analyzing $\mathbf{R}$ as if they were $\mathbf{S}$ matrices (Jöreskog and Sörbom, 1996a, p. 35). However, as Cudeck (1989) pointed out, statistical theory for the analysis of CFA structures has been most completely developed for applications to $\mathbf{S}$, and the analysis of $\mathbf{R}$ may be problematic in many situations. The analysis of $\mathbf{R}$ with ML may modify the model that is being analyzed, may product incorrect χ^2 and derived indices, and may give incorrect standard errors (Jöreskog et al., 2001). The only advantage of $\mathbf{R}$ is improving the interpretability of the solutions. When variables have quite different variances, may be useful analyzing $\mathbf{R}$ matrices. This is a common practice when fitting a regression models or using an Exploratory Factor Analysis (Cudeck, 1989, pp. 325–326). The general recommendation is to use ML estimation method with $\mathbf{S}$ and ULS with $\mathbf{R}$ (for a more detailed description, see Bollen, 1989, pp. 104–113). In addition, when a $\mathbf{S}$ matrix is analyzed, the completely standardized solution can be obtained, so it is no necessary to analyze $\mathbf{R}$ to obtain a more interpretable solution. Note that a $\mathbf{R}$ matrix is a completely standardized $\mathbf{S}$ matrix (Brown, 2015). A full description of how analyzing polychoric/tetrachoric matrices, with PRELIS/LISREL syntax examples, can be found in Yang-Wallentin et al. (2010).

Example 3 from Presentation 1 shows LISREL syntax (Jöreskog and Sörbom, 1996a) to estimate CFA models from the simulated dataset. A convergence criterion equal to 250 has been fixed in all the simulated conditions.

DATASET OVERVIEW AND OUTPUTS

The Supplementary Material includes a set of files with simulated raw-sample data from all the experimental conditions. The name of each file has two parts. The letter "N" denotes the sample size used to generate the data collected by each file (N200, N300, N400, and N500), and L020, L030 and L040 denote the simulated (λ_{ik}) magnitude (0.2, 0.3, and 0.4, respectively). There are 24 SPSS/Excel files with raw data, (e.g., for N = 200: N200L020, N200L030, N200L040). In addition, we present the estimated solutions conducting CFA with LISREL program from the proposed simulation study. An overview of all of this

information and outputs can be consulted in Presentation 2, and the results of each estimation process can be examined in two dataset:

- Continuous_data_INFORMATION.pdf.
- Continuous_Data_Normal_ML.
- Continuous_Data_Normal_ULS.

The last two types of files contain the parameter estimates, the standard errors and different goodness-of-fit measures. The analysis of **S** with ML generate unstandardized solutions (Brown, 2015). The unstandardized parameter estimates can be consulted in Unstandardized_parameters_ML files. To facilitate the direct comparison between ML and ULS estimators, ML files contain standardized solutions (default option in analysis of **R**). Note that LISREL provides standard errors for each estimated parameter in the unit of measurement of the indicators (Bollen, 1989; Jöreskog and Sörbom, 1996a). Therefore, to construct statistical tests (t-test) for estimated factor loadings (λ^*_{ik}) with the parameter estimates provided by ML method, the unstandardized values must be used.

Coefficient of congruence (see Equation 3) was computed as an index of factor similarity:

$$C_k = \frac{\Sigma_{i=1}^{p} \lambda^*_{ik} \lambda_{ik}}{\sqrt{(\Sigma_{i=1}^{p} \lambda^{*2}_{ik})(\Sigma_{i=1}^{p} \lambda^{2}_{ik})}} \qquad (3)$$

This coefficient computes the discrepancy between λ_{ik} and λ^*_{ik} for each indicator i of factor k, where p is the number of items per factor (p/k), and reflects a combined measure of good or poor parameter recovery of a given cluster of indicators. Lorenzo-Seva and Ten Berge (2006) have shown that congruence values in the range of 0.85–0.95 can be considered as "fair similarity" between λ_{ik} and λ^*_{ik}, and values higher than 0.95 as "good similarity".

Through the study shown in this work, the researchers can use some of the outputs attached in the Supplementary Material to evaluate different general issues related to the potential quality and accuracy of the CFA models. Two applications are described below:

Application 1: The researchers can explore the accuracy of parameter estimation, and the statistical relationship between the parameter recovery (C_k), the overall model fit (χ^2), and a high number of goodness-of-fit indices. Note that the χ^2 values of the output files depend on the method of estimation used to fit the model to the data (Jöreskog, 2004). **Table 1** shows a descriptive summary of this type of information obtained from continuous data ML file. First, we have filtered the dataset removing the improper solutions (non-convergent solutions and Heywood cases). As can be seen, the parameter recovery (C_k) improves as λ_{ik}, p/k, and N increase.

Table 1 shows that several of the estimated solutions present a certain degree of overestimation or underestimation of the parameters (λ^*_{ik}). A better understanding of the
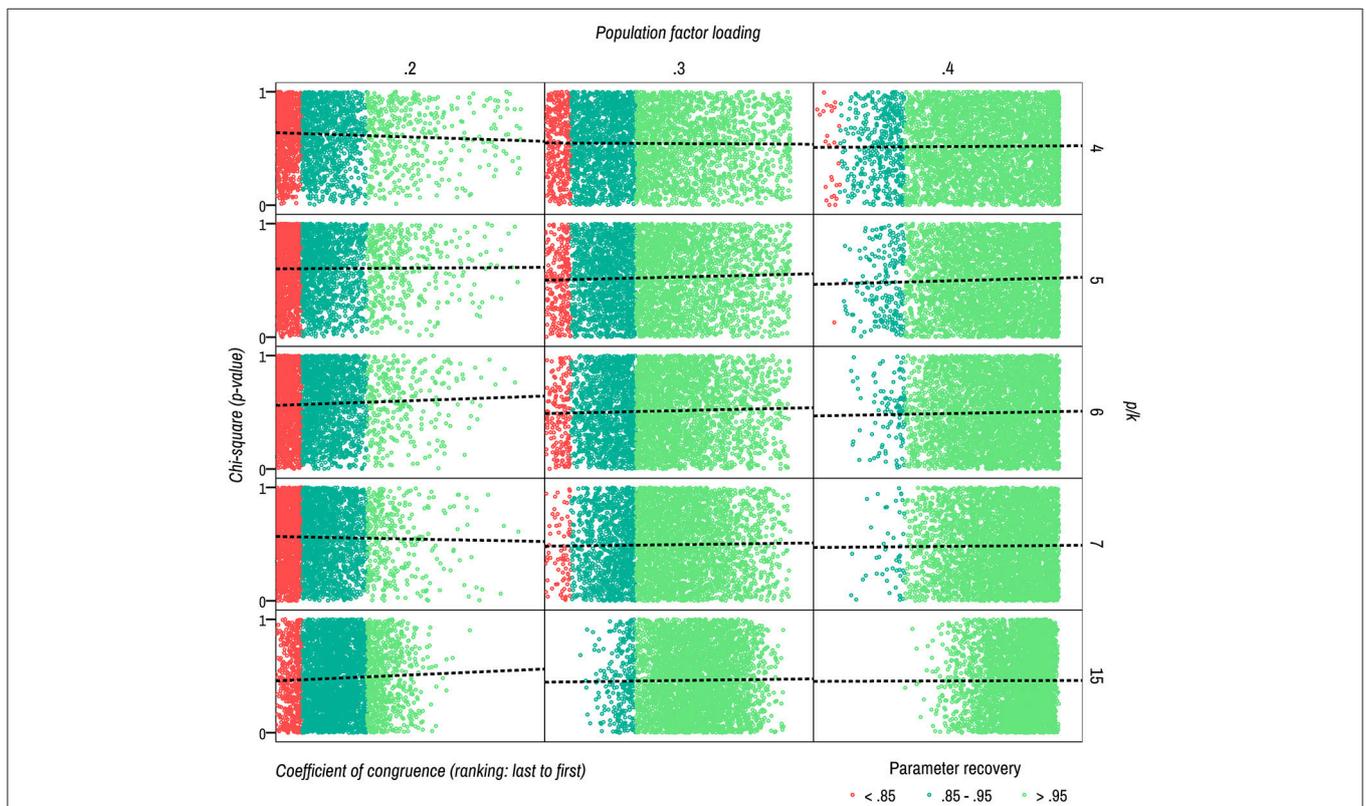
FIGURE 1 | Scatter plots for parameter recovery (x-axis) and overall fit (y-axis). ML estimation. Ranking of coefficient of congruence show the relative position of C_k values from all estimated solutions (ordered from solution with lowest C_k -last position- to higher C_k -first position-).

TABLE 1 | Descriptive summary of estimated CFA solutions (ML method): rate of proper solutions, parameter estimates (factor average), coefficient of congruence, RMSEA and CFI [%, Mean, (SD), min and max].

λ_{ik}		4–6 indicators		7 indicators		15 indicators	
		$N = 200–300$	$N = 400–500$	$N = 200–300$	$N = 400–500$	$N = 200–300$	$N = 400–500$
0.2	Proper solutions (%)	62.9%	72.5%	75.8%	88.7%	98.1%	100%
	Average λ^*_{ik}	0.24 (0.04)	0.22 (0.03)	0.22 (0.03)	0.21 (0.03)	0.20 (0.02)	0.20 (0.02)
	Average λ^*_{ik} (min-max)	0.14–0.38	0.10–0.35	0.13–0.32	0.12–0.30	0.13–0.28	0.11–0.26
	C_k	0.84 (0.09)	0.86 (0.09)	0.84 (0.09)	0.87 (0.08)	0.88 (0.06)	0.93 (0.03)
	RMSEA	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	CFI	0.70 (0.42)	0.82 (0.33)	0.78 (0.35)	0.86 (0.26)	0.85 (0.22)	0.93 (0.10)
0.3	Proper solutions (%)	93.0%	98.9%	99.1%	100%	100%	100%
	Average λ^*_{ik}	0.31 (0.04)	0.30 (0.03)	0.30 (0.03)	0.30 (0.02)	0.30 (0.02)	0.30 (0.02)
	Average λ^*_{ik} (min-max)	0.15–0.44	0.17–0.42	0.20–0.41	0.20–0.38	0.22–0.38	0.25–0.35
	C_k	0.92 (0.06)	0.96 (0.04)	0.94 (0.04)	0.97 (0.02)	0.97 (0.02)	0.98 (0.01)
	RMSEA	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	CFI	0.94 (0.15)	0.97 (0.08)	0.95 (0.09)	0.97 (0.05)	0.97 (0.05)	0.98 (0.02)
0.4	Proper solutions (%)	99.8%	100%	100%	100%	100%	100%
	Average λ^*_{ik}	0.40 (0.04)	0.40 (0.03)	0.40 (0.03)	0.40 (0.02)	0.40 (0.02)	0.40 (0.02)
	Average λ^*_{ik} (min-max)	0.25–0.52	0.28–0.50	0.28–0.50	0.32–0.47	0.33–0.48	0.35–0.45
	C_k	0.97 (0.02)	0.99 (0.01)	0.98 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.00)
	RMSEA	0.02 (0.03)	0.01 (0.02)	0.02 (0.02)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	CFI	0.98 (0.06)	0.99 (0.04)	0.98 (0.03)	0.99 (0.02)	0.99 (0.02)	0.99 (0.01)

conditions in which this lack of precision occurs is an important issue. Especially in the case of overestimation, the applied researchers may be too optimistic in the theoretical interpretation of the evaluated models. Moreover, **Table 1** also shows that the Root Mean Square Error (RMSEA) are not informative of this lack of accuracy (Heene et al., 2011). The Comparative Fit Index (CFI) is more informative in the most suboptimal conditions (i.e., $\lambda_{ik} = 0.2$). This situation may be contributing to confuse the theoretical interpretations of some applied researchers. Additionally, there are no relationship between accuracy of the parameter recovery (C_k) and χ^2 p-value, as can be seen in **Figure 1**.

Researchers and scholars can reuse these databases in order to explore in depth the conditions under which over-underestimation of the parameters may occur, or to evaluate the performance of the alternative goodness-of-fit indices. They can also reuse the provided syntax examples in order to extend the number and complexity of the experimental conditions.

Application 2: Each dataset provided in Supplementary Material has a "CASE" variable. This variable allows the raw data location of each CFA estimated solution collected in the two main output files, or a specific unstandardized solution for ML estimation, and it can be useful to review in depth the internal structure of certain estimated CFA models. For example, CASE = 90 ($\lambda_{ik} = 0.3$, $p/k = 4$, and $N = 200$) results in an improper solution when the ML estimation method is used. There are many situations that can

lead to an improper solution when a CFA model is fitted to the data (Brown, 2015). The condition of positive definiteness input matrix can be evaluated by Principal Component Analysis (PCA). If all eigenvalues are greater than zero, as in CASE = 90, the matrix is positive definite (Wothke, 1993). Then, we have explored the effect of outliers over estimation process. We have computed a new normally distributed random variable in the CASE = 90 raw dataset, that has been used as a criterion variable in a regression analysis with the X_1 to X_4 indicators as predictors. This analysis has helped us to identify outliers in the data (from Mahalanobis distances that are statistically significant at 0.05). Once these cases have been removed from raw data, CFA analysis produced an estimable-proper solution with $C_k = 0.863$.

CONCLUSION

Following open science philosophy (Munafò et al., 2017), in this paper we have shown how to estimate unifactorial models conducting CFA from a Monte Carlo simulation study with PRELIS and LISREL programs, which are some of the most used in Structural Equation Modeling (SEM) and CFA analysis. Additionally, we have made some indications about the estimation of CFA multidimensional models, with syntax examples, in order to facilitate the generalization of CFA analysis in more complex population structures. The applications presented are not intended to be an exhaustive evaluation of the results obtained, but rather an exposition of the possibilities of

the data generated and the estimation process implemented that may be of interest for academic and research purposes.

AUTHOR CONTRIBUTIONS

DO and JA developed the idea together. DO conducted the Monte Carlo simulation study with PRELIS/LISREL programs and collect all outputs into different dataset. JA supervised all the simulation process. DO drafted the manuscript and all authors approved the final version after discussing the intellectual content. All authors agreed to be accountable for all aspects of the work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg. 2018.00751/full#supplementary-material

Data Availability

In the Supplementary Material (https://figshare.com/s/ 18eb0e998150d39bc952) we have attached both the simulated data and the CFA results (parameter estimates, standard errors and goodness-of-fit measures).

REFERENCES

Bollen, K. A. (1989). *Structural Equations With Latent Variables*. New York, NY: John Wiley and Sons.

Brown, T. (2015). *Confirmatory Factor Analysis for Applied Research, 2nd Edn*. New York, NY: Guilford Press.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychol. Bull.* 105, 317–327. doi: 10.1037/0033-2909.105.2.317

Harrison, R. L. (2010). Introduction To Monte Carlo Simulation. *AIP Conf. Proc.* 1204, 17–21. doi: 10.1063/1.3295638

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., and Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. *Psychol. Methods* 16, 319–336. doi: 10.1037/a0024917

Jackson, D. L., Gillaspy, Jr, J. A., and Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychol. Methods* 14, 6–23. doi: 10.1037/a0014694

Jöreskog, K. G. (2004). *On Chi-Squares for the Independence Model and Fit Measures in LISREL, 1–10*. Available online at: http://www.ssicentral.com/ lisrel/techdocs/ftb.pdf

Jöreskog, K. G., and Sörbom, D. (1996a). *LISREL 8: User's Reference Guide*. Lincolnwood, IL: Scientific Software International.

Jöreskog, K. G., and Sörbom, D. (1996b). *PRELIS 2: User's Reference Guide*. Lincolnwood, IL: Scientific Software International.

Jöreskog, K. G., Sörbom, D., Du Toit, S. H. C., and Du Toit, M. (2001). *LISREL 8: New Statistical Features*. Lincolnwood, IL: Scientific Software International.

Kelley, K., and Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: evaluation of methods, recommendations, and software for composite measures. *Psychol. Methods* 21, 69–92. doi: 10.1037/a0040086

Lorenzo-Seva, U., and Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology* 2, 57–64. doi: 10.1027/1614-2241.2.2.57

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nat. Hum. Behav.* 1:21. doi: 10.1038/s41562-016-0021

Wothke, W. (1993). "Nonpositive definite matrices in structural modeling," in *Testing Structural Equation Models*, eds K. A. Bollen and J. S. Long (Newbury Park, CA: Sage), 256–293.

Yang-Wallentin, F., Jöreskog, K. G., and Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Struct. Equ. Modeling* 17, 392–423. doi: 10.1080/10705511.2010.489003