Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis

Ralph O Mueller, University of Hartford, West Hartford, CT, USA
Gregory R Hancock, University of Maryland, College Park, MD, USA

This article is a revision of the previous edition article by R.O. Mueller, volume 8, pp. 5239–5244, © 2001, Elsevier Ltd.

Abstract

Confirmatory factor analysis (CFA) is a quantitative data analysis method that belongs to the family of structural equation modeling (SEM) techniques. CFA allows for the assessment of fit between observed data and an a priori conceptualized, theoretically grounded model that specifies the hypothesized causal relations between latent factors and their observed indicator (manifest) variables. Because population-level equivalence between data and model cannot be shown with sample data, CFA should be viewed as a mainly disconfirmatory technique. That is, CFA facilitates the statistical rejection – or, at best, a tentative retention – of a specific theory regarding the factor(s) responsible for the observed relations in the data. If, on the other hand, the investigator's intentions are a mostly ungrounded exploration of relations suggested by the data, classical exploratory factor analysis is the more appropriate approach. In this article, typical steps in a CFA are introduced theoretically and via example: from model specification and identification, to parameter estimation, data-model fit assessment, and potential model modification. Didactic references are provided for a more in-depth study of CFA and SEM techniques in the social and behavioral sciences.

Overview

The term *factor analysis* describes a host of methods, all of which have the purpose of facilitating a better understanding of the unobserved variables (factors) that underlie a set of directly measurable and observed variables (for a nontechnical overview, see Bandalos and Finney, 2010 and references therein). These factors are often believed to represent constructs, psychological or otherwise, that have a direct bearing on the measured variables; as such, they are assumed to motivate (and in turn be inferable from) the pattern of covariances (unstandardized correlations) among those observed variables. In the late 1960s, works by Karl Jöreskog (e.g., 1966, 1967) articulated a method for *confirmatory factor analysis* (CFA), an application of normal theory maximum likelihood estimation to factor models with specific theoretical latent structures. Such structures could include the a priori specification of the number of factors, their orthogonality or obliquity, and which variables had zero and nonzero relations with those factors. This distinguishes CFA from well-known *exploratory factor analysis* (for classic introductions, see Gorsuch, 1983; Mulaik, 1972) wherein the number and nature of the factors emerge from the observed variables' data through a mathematical algorithm, largely blind to any substantive theory. Most crucial in Jöreskog's CFA work was the provision for a formal statistical χ^2-test of the fit between the pattern of relations among the measured variables and the theorized factor model, thereby facilitating the disconfirmation or tentative confirmation of a hypothesized factor model. Soon after, Jöreskog and others put forth a more general framework for the integration of measured and latent variables into causal networks, serving as the foundation for what is often known as *structural equation modeling* (SEM). CFA, which may be considered a special case within the more general SEM framework, is the focus of the current article.

CFA Model Specification, Identification, and Parameter Estimation

Suppose an educational researcher wishes to investigate the possibility of a low positive relation between reading and mathematics ability for fifth grade students, which is measured by standardized tests such as the Stanford Achievement Test or the Iowa Test of Basic Skills. The model shown in Figure 1 might be hypothesized. Measured variables X_1 through X_6, shown in rectangles, are believed to be caused by the latent factors ξ_1 and ξ_2, shown in ellipses. Here, ξ_1 and ξ_2 represent

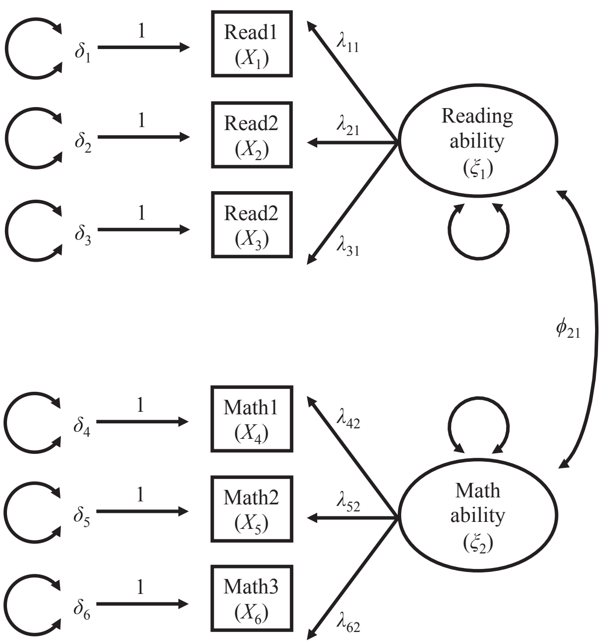

Figure 1 Hypothetical CFA model of reading and mathematics ability.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25009-5

true latent (unobserved) reading and mathematics ability, respectively, with X_1 through X_3 being standardized reading test measures (Read1 through Read3) and X_4 through X_6 being standardized mathematics test measures (Math1 through Math3). Table 1 includes a simulated variance/covariance matrix for the observed variables X_1 through X_6 based on test data from $n = 1200$ fifth graders. Our example's focus is the noncausal covariance between reading and mathematics ability, ϕ_{21}. In SEM path diagrams, a covariance is indicated by a two-headed arrow connecting the two constructs, and – because a variance is a covariance of a variable (observed or latent) with itself – it, too, is depicted by a two-headed arrow, in this case from the variable to itself.

A factor's hypothesized causal impact on its measured indicator variables, its *loading*, is symbolized by an arrow from the factor to the variable with magnitude λ_{ij}, where i denotes the observed variable and j denotes the latent factor. Note that such a model explicitly posits the factors as causing the variables, rather than the variables causing the factors; the latter type of model, in which the factor is characterized as *emergent* rather than *latent*, is much less common and beyond the scope of this article (but see Kline, 2013). In many cases, there is no arrow from a factor to a variable, such as from ξ_1 to X_4; this implies that reading ability has no theoretical causal bearing on the Math1 variable. Finally, to the extent that the factors do not perfectly explain each variable, a residual term, δ_i, is included as an influential contributor (with its variance shown by a two-headed arrow from δ_i to itself). This residual might consist of variable-specific measurement error as well as other influences. Thus, each observed variable is the sum of two parts, that attributable to the latent factor(s) and a residual part unique to the variable.

The causal relations of the hypothesized model shown in Figure 1 may be expressed as a system of six regression-like *structural equations*:

$$X_1 = \lambda_{11}\xi_1 + 0\xi_2 + \delta_1 \quad [1]$$

$$X_2 = \lambda_{21}\xi_1 + 0\xi_2 + \delta_2 \quad [2]$$

$$X_3 = \lambda_{31}\xi_1 + 0\xi_2 + \delta_3 \quad [3]$$

$$X_4 = 0\xi_1 + \lambda_{42}\xi_2 + \delta_4 \quad [4]$$

$$X_5 = 0\xi_1 + \lambda_{52}\xi_2 + \delta_5 \quad [5]$$

$$X_6 = 0\xi_1 + \lambda_{62}\xi_2 + \delta_6 \quad [6]$$

Equivalently, these equations can be represented in matrix form, as in

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \quad [7]$$

That is,

$$\mathbf{X} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \quad [8]$$

where $\mathbf{X}$ is a column vector of observed variables, Λ is a matrix of factor loadings, $\boldsymbol{\xi}$ is a vector of latent constructs, and $\boldsymbol{\delta}$ is a column vector of residuals.

The implication of Figure 1 and the accompanying structural equations is that the population variance/covariance matrix for the X variables, $\mathbf{\Sigma}$, is a function of (1) the λ_{ij} loadings in matrix Λ, (2) the variances and covariance of the latent factors in a matrix $\mathbf{\Phi}$, and (3) the variances and covariances among the residuals in a matrix $\mathbf{\Theta}_{\boldsymbol{\delta}}$ (note that in Figure 1 all residual covariances are zero as implied by the absence of two-headed arrows between the δ terms). More specifically, if all model parameters (loadings, variances, and covariances) are contained in a single column vector $\boldsymbol{\theta}$, the population variance/covariance matrix of the observed variables that is implied by the model and its parameters, $\mathbf{\Sigma}(\boldsymbol{\theta})$, is given by

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta}_{\boldsymbol{\delta}} \quad [9]$$

A vector of parameter estimates, $\widehat{\boldsymbol{\theta}}$, can be derived so that the model-implied variance/covariance matrix $\mathbf{\Sigma}(\widehat{\boldsymbol{\theta}})$ is as similar as possible to the observed variance/covariance matrix, $\mathbf{S}$, provided that model identification has first been ensured. To that end, each parameter in a model must be expressible as a function of the variances and covariances of the observed variables. When a system of such relations can be uniquely

Table 1 Simulated data and selected parameter estimates for the reading and mathematics ability model in Figure 1

	X_1	X_2	X_3	X_4	X_5	X_6
Variance/covariance matrix (n = 1200)						
	129.96					
	79.75	192.65				
	694.20	871.11	12038.48			
	307.03	391.75	3402.04	9876.38		
	230.99	415.74	2476.89	4815.24	12126.41	
	37.85	53.36	416.71	740.98	656.84	135.722
Standardized factor loadings and indicator reliability estimates						
ξ_1	0.70*	0.73*	0.79*	0	0	0
ξ_2	0	0	0	0.75*	0.60*	0.85*
R^2	0.49	0.53	0.62	0.56	0.35	0.73
Data-model fit indices						
$\chi^2 = 16.98$, $df = 8$ $p = 0.030$		CFI = 0.996	SRMR = 0.016		RMSEA = 0.031	

Note. *$p < 0.05$.

solved for the unknown parameters, the model is *just-identified*. When multiple such expressions exist for one or more parameters, the model is *overidentified* and, in that case, a best-fit (although not unique) estimate for each parameter is derived. If, however, at least one parameter cannot be expressed as a function of the observed variables' variances and covariances, the model is *underidentified*, and some or all parameters cannot be estimated on the basis of the data alone. This underidentification might be the result of the researcher attempting to impose a model that is too complex relative to the number of variances and covariances of the observed variables. Additionally, empirical underidentification might arise when unfortunate estimates for select parameters (e.g., values of zero for factor covariances) render subsets of model parameters inestimable. Fortunately, in most CFA applications, it suffices to ensure that (1) the total number of parameters to be estimated, t, does not exceed the number of unique variances and covariances of the observed variables, u, and (2) each latent factor has an assigned unit of measurement. To accomplish the latter condition for the model in **Figure 1**, we set the factor variances to unity (alternatively, for each of the two factors we could have specified one of the factor loadings to equal unity, thereby setting each factor's units equal to those of the observed variable). The model in **Figure 1** is overidentified with $u = 6(7)/2 = 21$ nonredundant observed variances and covariances and $t = 13$ parameters to be estimated: six factor loadings, one covariance between the two latent factors, and six variances of the residual terms associated with the observed variables.

Given that a model is just- or (preferably) overidentified, sample estimates can be obtained through a variety of estimation methods. These include maximum likelihood and generalized least squares, both of which assume multivariate normality and are asymptotically equivalent, as well as asymptotically distribution-free estimation methods that generally require a substantially larger sample size. These methods iteratively minimize a function of the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$, where $\mathbf{S}$ is the unrestricted variance/covariance matrix of the observed X variables and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ is the model-implied variance/covariance matrix reproduced from the iteratively changing parameter estimates. The standardized maximum likelihood estimates of key parameters in the reading and mathematics ability model are presented in **Table 1**. Before focusing on the estimate of the example's main parameter of interest (ϕ_{21}), we should consider whether or not there is any evidence suggesting data-model misfit, any statistical – and theoretically justifiable – rationale for modifying the hypothesized model, or any indication of factor unreliability.

Data-Model Fit Assessment and Model Modification

One of the advantages of CFA is the ability to assess the quality of the fit of the data to the model. A multitude of measures exists that assist the researcher in deciding whether to reject or tentatively retain an a priori specified overidentified model (see Marsh et al., 1988; Tanaka, 1993). In general, measures to assess the fit between the variances and covariances observed in the data and those implied by the model can be classified into three categories: absolute, parsimonious, and incremental. *Absolute* fit indices are those that improve as the overall discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ decreases. Examples of such measures include the model χ^2 statistic that tests the stringent null hypothesis H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, the standardized root mean-square residual (SRMR) that roughly assesses the average standardized discrepancy between observed and model-implied variances and covariances, and the goodness-of-fit index (Jöreskog and Sörbom, 1996) designed to evaluate the amount of observed variance/covariance information that can be accounted for by the model.

Parsimonious fit indices take into account not just the overall absolute fit, but also the degree of model complexity required to achieve that fit. Indices such as the adjusted goodness-of-fit index (Jöreskog and Sörbom, 1996), the Akaike Information Criterion (Akaike, 1974), and the root mean-square error of approximation (RMSEA, Steiger and Lind, 1980) indicate greatest data-model fit when data have reasonable absolute fit and models are relatively simple. Finally, *incremental* fit indices such as the normed fit index (Bentler and Bonett, 1980) and the comparative fit index (CFI, Bentler, 1990) gauge the data-model fit of a hypothesized model relative to that of a more restrictive baseline model with fewer parameters.

The three types of fit indices together help the researcher to converge upon a decision regarding the CFA model's acceptability. In our example, the statistically significant maximum likelihood $\chi^2 = 16.98$ ($df = u - t = 8$, $p < 0.05$) indicates that the observed variance/covariance matrix would occur rarely if our model correctly depicted the true population relations. This absolute fit statistic, however, is notoriously sensitive to very small and theoretically trivial model mis-specifications (e.g., slight amounts of error covariance) under large sample conditions. As such, other fit indices are generally preferred for model evaluation. According to Hu and Bentler (1999), for example, CFI values of 0.96 or greater together with SRMR values less than 0.09 (or with RMSEA values less than 0.06) point to acceptable data-model fit. The indices in **Table 1** suggest no appreciable data-model inconsistency given the Hu and Bentler recommendations.

After the data-model fit has been assessed, a decision about that model's worth must be reached. Acceptable fit indices usually lead to the conclusion that no present evidence exists warranting a rejection of the model or the theory underlying it. This is not to say that the model and theory have been confirmed, much less proven as correct; rather, the current factor model remains as one of possibly many that satisfactorily explain the relations among the observed variables (see Hershberger and Marcoulides, 2013). On the other hand, when fit indices indicate a potential data-model misfit, one might be reluctant to dismiss the model entirely. Instead, attempts are often made to modify the model post hoc so that acceptable fit indices can be obtained. Such modifications could include the addition of cross-loadings, allowing a given variable to load on multiple factors, or of error covariances in which variables' residuals are allowed to covary in order to reflect some potential variable relation above and beyond that motivated by the factor structure itself.

Most CFA software packages will facilitate such model improvement by providing modification indices (Lagrange multiplier tests) indicating what changes in the model could reap the greatest increase in absolute fit, that is, decrease in the model χ^2 statistic. While such indices constitute a potentially useful tool for remedying incorrectly specified models, it seems imperative to warn against an atheoretical hunt for the model with the best fit. Many alternative models exist that can explain the observed data equally well; hence, attempted modifications must be based on a sound understanding of the specific theory underlying the model. Furthermore, when modifications and reanalyzes of the data are based solely on data-model misfit information, subsequent fit results might be due largely to chance rather than true improvements to the model. Modified structures therefore should be cross-validated with an independent sample whenever possible. If a new sample is not available but the initial sample is large enough, one can randomly split the sample into calibration and validation subsamples and compute Cudeck and Browne's (1983) cross-validation index. When the initial sample is too small, Browne and Cudeck (1989) also offered a single sample alternative, an estimate of the expected value of the cross-validation index. From the analysis of the model in **Figure 1**, none of the modification indices suggested changes to the model that would result in a significant improvement in data-model fit (i.e., a significant decrease in the model χ^2 statistic); thus, we did not report them in **Table 1**.

Before drawing conclusions regarding the relation between reading and mathematics ability, the question of the quality (i.e., reliability) of the factors should be addressed. Traditionally, this has been accomplished (1) by focusing on the reliability of scores from individual indicator variables, R^2 (i.e., the proportion of variability in an observed variable that can be accounted for by the underlying factor), or (2) by assessing the reliability of scores from linear composites of the indicator variables (e.g., Miller, 1995; Raykov, 1997). Alternatively, Hancock and Mueller (2001) suggested a measure (coefficient H) computable most easily from the standardized factor loadings that can be used to assess the reliability (i.e., replicability) of a latent construct itself as reflected by scores from its multiple observed indicator variables. One of coefficient H's advantages over traditional construct reliability measures is that it is never less than the best indicator variable's reliability (R^2), thereby drawing information from all indicators in a manner commensurate with their own ability to reflect the construct. For the current example, $\widehat{H} = 0.79$ and $\widehat{H} = 0.82$ for the reading and mathematics ability constructs, respectively, while the R^2 values for the respective factors' indicator variables in **Table 1** range between 0.49–0.62 and 0.35–0.73. Thus, the two factors exhibit reasonable and satisfactory levels of construct reliability, given that about 80% of their variance is explainable by their respective indicators. **Table 1** also lists the standardized factor loadings, that is, the standardized versions of the λ_{ij} parameters in the $\mathbf{\Lambda}$ matrix that indicate the strength and direction of the a priori specified causal influences of the latent factors on the observed variables. All factor loadings are positive and statistically significantly different from zero ($p < 0.05$).

Finally, our primary research question can be addressed. The CFA estimate of the correlation between the two latent constructs reading and mathematics ability is 0.51 and statistically significantly different from zero ($p < 0.05$). This estimate indicates a low to moderate positive association between the two constructs of interest, as hypothesized.

Conclusion

CFA has become established as an important analytical tool for many areas of the social and behavioral sciences. It belongs to the family of SEM techniques that allow for the investigation of causal relations among latent and observed variables in a priori specified, theory-derived models. The main advantage of CFA lies in its ability to aid researchers in bridging the common gap between theory and observation. For example, an instrument might be developed by creating multiple items for each of several specific theoretical constructs (**Figure 1**). Instead of analyzing data with an exploratory factor analysis (where each item is free to load on each factor) and potentially facing a solution inconsistent with initial theory, a CFA can give the investigator valuable information regarding the fit of the data to the specific, theory-derived measurement model (where items load only on the factors they were designed to measure), and point to the potential weakness of specific items. CFA is best understood as a process, from model conceptualization, identification, and parameter estimation, to data-model fit assessment and potential model modification. As opposed to exploratory methods, CFA's strength lies in its disconfirmatory nature: models or theories can be rejected, but results might also point toward potential modifications to be investigated in subsequent analyses.

A recent treatment of CFA for applied researchers is Brown (2009). Popular textbooks on more general SEM – those that include many CFA examples from disciplines covered in this encyclopedia and that utilize commonly available software (i.e., AMOS, EQS, LISREL, and MPlus) – are Bollen (1989), Byrne (1998, 2006, 2010, 2012), Kline (2011), Loehlin (2004), Raykov and Marcoulides (2006), and Schumacker and Lomax (2010). Social science journals that publish many CFA and SEM applications and methodological developments include *Educational and Psychological Measurement*, *Multivariate Behavioral Research*, *Journal of Experimental Education*, *Psychological Methods*, *Sociological Methodology*, *Sociological Methods and Research*, and *Structural Equation Modeling: A Multidisciplinary Journal*.

See also: Big Five Factor Model, Theory and Structure; Emotions and Aging; Factor Analysis and Latent Variable Models in Personality Psychology; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Five Factor Model of Personality, Universality of.

Bibliography

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Bandalos, D.L., Finney, S.J., 2010. Factor analysis: exploratory and confirmatory. In: Hancock, G.R., Mueller, R.O. (Eds.), The Reviewer's Guide to Quantitative Methods in the Social Sciences. Routledge, New York, NY, pp. 93–114.

Bentler, P.M., 1990. Comparative fit indexes in structural models. Psychological Bulletin 107, 238–246.

Bentler, P.M., Bonett, D.G., 1980. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin 88, 588–606.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York, NY.

Brown, T.A., 2009. Confirmatory Factor Analysis for Applied Research. Guilford Press, New York, NY.

Browne, M.W., Cudeck, R., 1989. Single sample cross-validation indexes for covariance structures. Multivariate Behavioral Research 24, 445–455.
Byrne, B.M., 1998. Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS. Lawrence Erlbaum, Mahwah, NJ.
Byrne, B.M., 2006. Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming, second ed. Lawrence Erlbaum, Mahwah, NJ.
Byrne, B.M., 2010. Structural Equation Modeling with AMOS, second ed. Routledge, New York, NY.
Byrne, B.M., 2012. Structural Equation Modeling with Mplus. Taylor & Francis, New York, NY.
Cudeck, R., Browne, M.W., 1983. Cross-validation of covariance structures. Multivariate Behavioral Research 18, 147–167.
Gorsuch, R.L., 1983. Factor Analysis, second ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
Hancock, G.R., Mueller, R.O., 2001. Rethinking construct reliability within latent variable systems. In: Cudeck, R., du Toit, S., Sörbom, D. (Eds.), Structural Equation Modeling: Present and Future—A Festschrift in Honor of Karl Jöreskog. Scientific Software International, Lincolnwood, IL, pp. 195–216.
Hershberger, S.L., Marcoulides, G.A., 2013. The problem of equivalent models. In: Hancock, G.R., Mueller, R.O. (Eds.), Structural Equation Modeling: A Second Course, second ed. Information Age Publishing, Charlotte, NC, pp. 3–39.
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, 1–55.
Jöreskog, K.G., 1966. Testing a simple structure hypothesis in factor analysis. Psychometrika 31, 165–178.
Jöreskog, K.G., 1967. Some contributions to maximum likelihood factor analysis. Psychometrika 32, 443–482.
Jöreskog, K.G., Sörbom, D., 1996. LISREL 8. Scientific Software International, Chicago.
Kline, R.B., 2011. Principles and Practice of Structural Equation Modeling, third ed. The Guilford Press, New York, NY.
Kline, R.B., 2013. Reverse arrow dynamics: Feedback loops and formative measurement. In: Hancock, G.R., Mueller, R.O. (Eds.), Structural Equation Modeling: A Second Course, second ed. Information Age Publishing, Charlotte, NC, pp. 41–79.
Loehlin, J.C., 2004. Latent Variable Models, fourth ed. Lawrence Erlbaum, Hillsdale, NJ.
Marsh, H.W., Balla, J.R., McDonald, R.P., 1988. Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size. Psychological Bulletin 103, 391–410.
Miller, M.B., 1995. Coefficient alpha: a basic introduction from the perspectives of classical test theory and structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal 2, 255–273.
Mulaik, S.A., 1972. The Foundations of Factor Analysis. McGraw-Hill, New York, NY.
Raykov, T., 1997. Estimation of composite reliability for congeneric measures. Applied Psychological Measurement 21, 173–184.
Raykov, T., Marcoulides, G.A., 2006. A First Course in Structural Equation Modeling, second ed. Lawrence Erlbaum, Mahwah, NJ.
Schumacker, R.E., Lomax, R.G., 2010. A Beginner's Guide to Structural Equation Modeling, third ed. Routledge, New York, NY.
Steiger, J.H., Lind, J.M., 1980. Statistically based tests for the number of common factors. In: Paper Presented at the Annual Meeting of the Psychometric Society. Iowa City, IA.
Tanaka, J.S., 1993. Multifaceted conceptions of fit in structural equation models. In: Bollen, K.A., Long, J.S. (Eds.), Testing Structural Equation Models. Sage, Newbury Park, CA, pp. 10–39.

Relevant Websites

www.ibm.com – AMOS Software.
www.mvsoft.com – EQS Software.
www.ssicentral.com – LISREL Software.
www.statmodel.com – MPlus Software.