Applications of Structural Equation Modeling in Social Sciences Research

Jackson de Carvalho, PhD
Assistant Professor
Division of Social Work, Behavioral & Political Science
Prairie View A&M University
United States of America

Felix O. Chima, PhD
Professor
Division of Social Work, Behavioral & Political Science
Prairie View A&M University
United States of America

Abstract

Structural equation modeling (SEM) is a comprehensive statistical modeling tool for analyzing multivariate data involving complex relationships between and among variables (Hoyle, 1995). SEM surpasses traditional regression models by including multiple independent and dependent variables to test associated hypothesizes about relationships among observed and latent variables. SEM explain why results occur while reducing misleading results by submitting all variables in the model to measurement error or uncontrolled variation of the measured variables. The purpose of this article is to provide basic knowledge of structural equation modeling methodology for testing relationships between indicator variables and latent constructs where SEM is the analysis technique of the research statistical design. It is noteworthy, SEM provides a way to test the specified set of relationships among observed and latent variables as a whole, and allow theory testing even when experiments are not possible. Consequently, these methodological approaches have become ubiquitous in the scientific research process of all disciplines.

Key Words: Structural equation modeling, methodology, multivariate analysis and research

Introduction

According to Byrne (2010), Structural Equation Modeling (SEM) is a powerful collection of multivariate analysis techniques, which specifies the relationships between variables through the use of two main sets of equations: Measurement equations and structural equations. Measurement equations test the accuracy of proposed measurements by assessing relationships between latent variables and their respective indicators. The structural equations drive the assessment of the hypothesized relationships between the latent variables, which allow testing the statistical hypotheses for the study. Additionally, SEM considers the modeling of interactions, nonlinearities, correlated independents, measurement error, correlated error terms, and multiple latent independents each measured by multiple indicators.

Given that SEM is heavily infused with jargon, particularly regarding the types of variables hypothesized in the model a preliminary definition of terms allow for a more clear explanation of the findings. Some of the common terminologies used in SEM include:

- **Exogenous Variables -** Variables that are not influenced by other variables in the model.
- **Endogenous Variables -** Variable that is caused by other variables in the model
- **Indicator Variables -** Variables that are directly observed and measured (also known as manifest variables in some circles).
- **Latent Variables -** Variables that are not directly measured
- **Measurement Model -** This is a part of the entire structural equation model diagram hypothesized for the study including all observations that load onto the latent variable, their relationships, variances, and errors.

- **Structural Model -** This is a part of the total hypothesized structural equation model diagram, which includes both latent and indicator variables
- **Structural Equation Model -** This model combines the structural model and the measurement model, which includes everything that has been measured and observed among the variables examined.

The difference between SEM and other conventional methods of statistical analysis is accentuated by significantly distinct characteristics. For example, the basic statistic in SEM is the covariance. It is worth to note that covariance statistics convey more information than a correlation (Hu & Bentler, 1999). While conventional regression analysis attempt to minimize differences between observed and expected individual cases, SEM aims to minimize differences between observed and expected covariance matrices. In other words, SEM, based on the covariance statistic, attempts "to understand patterns of correlations among a set of variables and to explain as much of their variances" (Kline, 1998, pp. 10-11).

Unlike conventional analysis, SEM allows the inclusion of latent variables into the analyses and it is not limited to relationships among observed variables and constructs. It allows the study to measure any combination of relationships by examining a series of dependent relationships simultaneously while considering potential errors of measurement among all variables. SEM has several advantages over conventional analysis, including greater flexibility regarding assumptions (particularly allowing interpretation even in the face of multicollinearity). SEM allows the use of confirmatory factor analysis to reduce measurement error by testing multiple indicators per latent variable while offering superior model visualization through its graphical modeling interface (Hatcher, 2005; Joreskog, 1993; Kline, 2005).

Moreover, SEM has the appealing capacity of testing models overall rather than coefficients individually. It also has the ability to test models with multiple dependent variables, to include mediating variables and to model error terms for all indicator variables. Another attractive quality of SEM is that as it considers potential errors of measurement in all variables and when a hypothesized structural model shows model fit indices that are less than satisfactory, it allows *specification searches* to find better fitting models to the sample variance-covariance matrix (Hu & Bentler, 1999; Kline, 2005; Schumacker & Lomax, 2004).

Overall, the structural equation modeling process centers around two steps. First, it validates the measurement model in terms of assessing the relationship between hypothetic latent constructs and clusters of observed variables underlying each construct. Validation o the measurement model is often conducted by using Confirmatory Factor Analysis (CFA). The second step centers around fitting the structural model by measuring the significance of the relationship between latent variables, which is often accomplished through path analysis (Hoyle, 1995; Kaplan, 2000).

In order to apply SEM in estimating relationships among variables, several computer programs such as CALIS, EQS, AMOS and LISREL can be used. AMOS computer program is often selected due to its suitability for essentially all stages of data analysis (Byrne, 2010; Kline, 2005, Schumacker & Lomax, 2004).

Drawing the Hypothesized Diagram Model

The main purpose of drawing a hypothesized diagram model in SEM is "to find a model that not only fits the data well from a statistical point of view, but also has the property that every parameter of the model can be given a substantively meaningful interpretation" (Joreskog, 1993, p. 295). The main steps involved in developing a structural equation model include:

- **Model Specification:** This is the process of formally stating a model by determining which parameters are to be fixed or free.
- **Model Identification:** This is the idea of having at least one unique solution for each parameter estimate in the model from the observed data.
- **Model Estimation:** This is process in which start values of the free parameters are chosen in order to generate an estimated population covariance matrix $\sum$ (θ), from the model (Loehlin, 2004).
- **Testing Model Fit**: This is the process of evaluating a ***structural equation model*** with goodness-of-***fit*** indices
- **Model Manipulation:** This is the process of making model adjustments through **specification** searches.

The process of drawing hypothesized diagram model is done in software (i.e., AMOS) and requires compliance with some of the following basic rules:

- Latent variables are depicted with circles
- Indicator variables are represented with squares.
- Lines with arrows in one direction represent a hypothesized direct relationship between the two variables.
- A curved line with arrows in both directions shows a covariance between two variables.
- Only exogenous variables have covariance arrows.
- Endogenous variables should have a residual term. A residual term is depicted by a circle with the letter E written in it, which stands for error.
- The error term in the endogenous latent variable is called a disturbance and it is depicted by a circle with a D written in it (Byrne, 2010).
- Parameters are the variances, regression coefficients and covariances among variables
- Variances are indicated by a two-headed arrow with both ends of the same arrow pointing at the same variable.
- Regression coefficients are depicted along single-headed arrows indicating a hypothesized pathway between two variables.
- Covariances are represented by double-headed, curved arrows between two variables
  or error terms.

Measurement Model

The process of validating the measurement model through confirmatory factor analysis (CFA) allows assessment of the research questions by determining whether the observed variables are indeed good indicators of the latent variables. Therefore, separate confirmatory factor models should run for each set of observed variables hypothesized to indicate their respective latent variable. Subsequently, the observed variables should be diagrammed (in AMOS, for example) and linked to an SPSS data file to test if the indicator variables are acceptable in defining the latent variable (Byrne, 2010; Schumacker & Lomax, 2004).

In an effort to improve model fit, the researcher may include correlation of error covariance between indicator variables as specified. That is, including correlated measurement error in the model tests the possibility that common or correlated variables not included in the *model* may be responsible for correlations between indicator variables. The possibility of extraneous correlation between indicators can be ruled out if the fit of the model specifying uncorrelated error terms is as good as the model with correlated error specified (Kline, 2005; Hatcher, 2005).

The process of validating the measurement model requires testing each cluster of observed variables separately to fit the hypothesized CFA model. The statistical test uses the most popular procedures of evaluating the measurement model: Chi-square ($\chi2$), Goodness-of-Fit Index (GFI), and Percent Variance Explained. The percent variance explained should be calculated as the sum of the communalities (h^2) divided by the number of variables ($\Sigma h^2/m$) (Bentler, 1990).

Chi-square should be divided by degrees of freedom (chi-square/df) at the expected ***ratio*** of ***two chi-square*** variables ***divided*** by their respective ***degrees of freedom*** and model fit statistics should be close to the $p < .05$ level of significance. The GFI represents the overall degree of fit, which are the squared residuals. Values of .90 or above for the GFI indicate a good fit and values below 0.90 simply suggest that the model can be improved. The model fit statistics for the measurement models should be summarized and presented in a separate table. Testing of the confirmatory factor model, however, may well be a desirable validation stage preliminary to the main use of SEM to identify the causal relations among latent variables (Schumacker & Lomax, 2004; Browne & Cudeck, 1993).

Structural Model

Although, the measurement model embodies the relationships between measured variables and latent variables, the structural model represents the relationships between latent variables only and it must be inferred from measured variables. Measured variables are those that can be observed directly while latent variables are not. In an effort to answer the research questions and underlying hypotheses the SEM approach allows examination of structural relationships among theoretical factors by analyzing the structural model and the measurement model separately (Schumacker & Lomax, 2004).

In order for the structural relationships among theoretical factors to be analyzed in SEM and generate a sensible set of results, an adequate number of known correlations or covariances are required. This analysis process is accomplished primarily through path analysis with latent variables of the acceptable model. The statistics literature shows no consistent standards for what is considered an acceptable model; a lower chi-square to df ratios, however, indicates a better model fit as chi-square statistics is one of the most commonly used techniques to examine model fit. The SEM approach, however, focuses on evaluating the values of the Adjusted Goodness of Fit Index (Values above 0.90) and the desirable ratio of chi square to degrees of freedom to be well below two indicating that the model fit data well (Byrne, 2010).

If an initial model shows statistical indices that are not acceptable, specification searches are conducted where modification indices may suggest adding additional paths in the existing model. This process may be repeated until a final model shows acceptable fit statistics. The main use of SEM to identify the causal relations among latent variables requires each equation to be properly identified in a process known as model identification and estimation, which is preliminary to testing the significance of the structure coefficients and associated statistics (McDonald & Ho, 2002; Rubin & Babbie, 2010).

Model Identification and Estimation

The concept of identification refers to the idea of having at least one unique solution for each parameter estimate in the model. When models have only one possible solution for each parameter estimate they are known to be just-identified. Models with infinite number of solutions are known to be underidentified and models with more than one possible solution, but has one best or optimal solution for each parameter estimate are known to be overidentified (Byrne, 2001; Schumacker and Lomax, 2004).

The SEM approach requires the model to be overidentified, which means that the number of data points is greater than the number of parameters to be estimated. The over identification imposes restrictions on the model, which allows for a test of the hypotheses specified. The identification of a hypothetical model should follow the following steps of SEM: 1) determine input matrix and estimation method, (2) assess the identification of the model, (3) evaluate the model fit, and (4) re-specify the model and evaluate the fit of the revised model (Byrne, 2010).

In step one, the Maximum Likelihood method (ML) should be utilized for the proposed model. Maximum likelihood is the procedure of finding the value of one or more parameters for a given statistic, which makes the known likelihood (the hypothetical probability that an event which has already occurred would yield a specific outcome distribution) the maximum value of a set of elements. Considering the current set of observations, the method of maximum likelihood finds the parameters of the model that are most consistent with these observations. The parameters of the model are: (1) variances and covariances of latent variables, (2) direct effects (path coefficients) on the dependent variable, and (3) variances of the disturbances (residual errors).

In step two, assessment of the ability of the proposed model to generate unique solutions should be conducted as an effective model identification process allows estimate for all the parameters independently and for the model as a whole. In step three, the overall model fit (the goodness of fit between the hypothesized model and the sample data) is assessed with several goodness-of-fit indexes. Since chi-square statistics is one of the most commonly used techniques to examine overall model fit, it is noteworthy that a non-significant goodness-of-fit X^2 statistic is favored because it indicates that the implied covariance matrix is nearly identical to the observed data. If the estimated covariance matrix does not provide a reasonable and parsimonious explanation of the data then the model may be re-specified by changing model parameters (Bollen & Long, 2010; Kline, 2005; McDonald & Ho, 2002).

Lastly, an adjustment of the hypothesized model is conducted by examining the goodness-of-fit indices to improve the model based on theoretical justification as the model is re-specified. The estimated covariance matrix may or may not provide a reasonable and parsimonious explanation of the data, which may lead to the model being accepted or rejected. Thus, an adjustment and improvement of the model allows identification of data related problems and potential sources of poor fit. Furthermore, the adjustment process can provide new insights regarding the relationship between observed and latent variables (Bollen & Long, 2010).

Once the final model is specified through an over identification process, the next step is to test the hypothesized model statistically to determine the extent to which the proposed model is consistent with the sample data, which includes the fit of the model as a whole and the fit of individual parameters.

The hypothesized model should be tested by using the two most popular ways of evaluating model fit: The X² goodness-of-fit statistic and fix indices. This process includes an examination of the parameter estimates, standard errors and significance of the parameter estimates, squared multiple correlation coefficients for the equations, the fit statistics, standardized residuals and the modification indices (Schumacker & Lomax, 2004; Rubin & Babbie, 2010).

Due to Chi-square's sensitivity to sample size, it is not easy to gain a good sense of fit solely from the X² value. Thus, other indexes of model fit should be examined, including: GFI (Goodness-of-Fit Index), AGFI (Adjusted Goodness-of-Fit Indices), CFI (Comparative Fit Index), SRMR (Standardized Root Mean Squared Residual) RMR (root mean square residual) and RMSEA (Root Mean Square Error of Approximation).

If the Goodness of Fit Index (GFI) is larger than 0.9 it reflects a good overall degree of fit and values below 0.90 simply suggest that the model can be improved. On the other hand, AGFI is the Adjusted Goodness of Fit Index. It considers the degrees of freedom available for testing the model. Values above 0.90 are acceptable, indicating that the model fits the data well. SRMR is the Standardized Root Mean Squared Residual, and it is a standardized summary of the average covariance residuals. SRMR should be less than .10 (Kline, 2005; Hatcher, 2005; Hu & Bentler, 1999).

General Structural Equation Model

General structural equation models include unobservable exogenous and endogenous variables (also termed factors or latent variables) in addition to the disturbances (error terms). Assessment of model fit indexes leads to the analysis of the general structural equation model to explain the relationship among the latent variables defined by the confirmatory factor models or measurement models as correlations, means, and standard deviations among the latent variables should be reported.

Subsequently, predicted correlations or covariances are compared to the observed correlations or covariances and if fit statistics are poor the model should be respecified and modification indices should be followed. The final modified model with acceptable model fit statistics should be used for testing the hypotheses related to the statistical significance of the structure coefficient or path in the model. A careful assessment of the structure coefficient, standard error, t-value and ***p-value will indicate if the null hypothesis should be accepted or rejected.***

Limitations Regarding SEM

The confirmatory technique used in SEM requires a preliminary knowledge of all the relationships to be specified in the model. It is crucial to know the number of parameters to be estimated – including covariances, path coefficients, and variances before beginning the data analysis. Additionally, tests of model fit are sensitive to sample size particularly when considering that SEM is often used to analyze complex relationships between multivariate data, which leads to the rejection of models even with trivial misspecifications. Trivial misspecifications refers to the size of the standardized residuals, which can be directly examined and determined if they are in fact trivial, but this is rarely done (Kline, 2005).

Furthermore, the underlying theory of SEM is fairly technical, and the modeling process can be frustrating and complicated. Consequently, this technique is often misused as researchers tend to develop a "fit index tunnel vision" (Kline, 2005, p. 321). Instead of considering multiple fit indices, researchers tend to ignore the test of fit and the residuals and consider only indices such as the CFI. It is noteworthy that residuals are the most informative set of fit indices as they point directly to the location and the size of the discrepancy between the covariances. Thus, fit index tunnel vision can lead researchers to erroneously avoid model modification due to excessively considering high CFI.

According to Kline (2005), SEM is a large sample technique. Conclusions extrapolated from a model based on a small sample size is unreliable as parameter estimation (variances, regression coefficients and covariances among variables) is often done by Maximum Likelihood (ML), which assumes normality among the indicator variables. If a large model is tested with a small sample size (less than 10 observations per variable), estimation problems are likely to occur. Unfortunately, most sample sizes in social sciences research are comprised of less than a few hundred cases, which challenges the assumption of maximum likelihood estimation as it is based on large sample sizes. In addition, data are seldom multivariate normal. Consequently, the mishmash of small sample sizes and nonnormal data can lead to estimation problems and unreliable results.

Despite of limitations, SEM is still a very powerful analytical tool with many advantages over other techniques. For instance SEM has the outstanding abilities of simultaneously using several indicator variables to define each construct in the model, which leads to more validity of the measurement model. In addition, SEM allows evaluation of relationships between constructs without random error, which distinguish SEM from other simpler, relational modeling techniques. Besides, social sciences research often assesses complex associations between and among variables, groups and conditions. SEM allows to model and test clusters of complex hypotheses simultaneously while assessing mean structures and group comparisons. Using other analysis technique, different than SEM, would require a number of separate steps.

References

Babbie, E. (2006). *The practice of social research (*11th ed.). Belmont, CA: Wadsworth.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34(2)* 181-197.

Byrne, B. M. (2010). Structural equation modeling with AMOS, (2nd ed.). New York: Routledge.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A.

Bollen & J S. Long (2010), *Testing structural equation models* (pp. 136-162). Newbury park, CA: Sage.

Hatcher, L. (2005). *A Step-By-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling.* Cary, NC: SAS Institute Inc.

Hoyle, R. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural Equation modeling: Concepts, issues, and applications* (pp. 1-15). Thousand Oaks CA: Sage.

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

Joreskog, K. (1993). Testing structural equation models. In K. A. Bollen & J. S. Logn (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage Publications.

Kline, R. (1998). *Principles and practice of structural equation modeling.* NY: Guilford Press.

Kline, R. (2005). *Principles and practices of structural equation modeling* (2n ed.). New York: Guilford Press.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural analysis* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

McDonald, R., & Ho, M. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64-82.

Pedhazur, E. & Pedhazur, L. (1991). *Measurement, Design and Analysis*. Lawrence Eblaum Associaltes, Publishers: Hillsdale, New Jersey.

Rubin, A. & Babbie, E. (2010*). Research methods for social* work. Belmont, CA: Wadsworth/Thomson Press.

Schumacker, R. Lomax, R. (2004). *A Beginner's Guide to Structural Equation Modeling 2nd Ed.* Mahwah, NJ: Lawrence Erlbaum