Asian Social Science; Vol. 11, No. 4; 2015
ISSN 1911-2017 E-ISSN 1911-2025
Published by Canadian Center of Science and Education

Use of Structural Equation Modeling in Social Science Research

Wali Rahman[1], Fayaz Ali Shah[2] & Amran Rasli[2]

[1] Serhad University, Peshawar, Pakistan

[2] Faculty of Management, Universiti Tecknologi, Malaysia

Correspondence: Fayaz Ali Shah, Faculty of Management, Universiti Tecknologi, Malaysia. E-mail: akhoon47@yahoo.com

Received: July 27, 2014 Accepted: January 7, 2015 Online Published: January 14, 2015

doi:10.5539/ass.v11n4p371 URL: http://dx.doi.org/10.5539/ass.v11n4p371

Abstract

A researcher mostly needs some statistical technique for the interpretation of the data at hand. This choice depends on the nature of the data and the researcher's own understanding and preferences of the available techniques. Structural Equation Modeling (SEM) is one among those techniques. The purpose of the present study is to present some basic aspects this powerful interdependence technique with and analysis of the most common issues of SEM. This paper will present a case as to how SEM excels other statistical techniques. Literature reveals that SEM is one of the most favored statistical techniques among the social science researchers and has been found to be better than other multivariate techniques including multiple regression analysis in examining series of dependence relationships simultaneously. However, it has been felt that the use of SEM in social research is equal to naught. Side by side there hardly exists any published review that systematically describes and critique the use of SEM. The present research is an endeavor to fill that gap. The study contributes to literature on SEM specifically and provides more holistic view of SEM for researchers to use SEM more effectively.

Keywords: SEM, multiple regression analysis, path analysis, management/social sciences

1. Introduction

According to Cheng (2001) Structural Equation Modeling (SEM) has been one of the most popular statistical techniques across various disciplines in the quantitative social sciences. This technique has got popularity because of the sophistication of its underlying theory and its potential for addressing important substantive questions (Kaplan, 2000). One of the exclusive features of this technique is its very potential of handling complex relationships among variables, where some variables can be hypothetical or unobserved. It is model based and a researcher can try one or more competing models. This analytics of this technique help a research as to which models fit, where there are redundancies, and can also help pinpoint what particular model aspect are in conflict with the data. It is a sort of a combination of multiple regression and factor analysis with some additional benefits over these techniques including an effective way to deal with multicollinearity, and methods for taking into account the unreliability of response data (Bacon, 1997). It is a statistical technique "that seeks to represent hypotheses about the means, variances, and covariances of observed data in terms of a smaller number of structural' parameters defined by a hypothesized underlying conceptual or theoretical model. It is a framework for using statistical methods to ask complex questions of data. Researchers employ different software for this statistical technique, provide a number of goodness of fit statistics for model testing thus reducing reliance on only a single statistic which in the past had mainly been the chi-square test. SEM technique appears to serve an increasingly important role in developing knowledge for the social work profession. This technique is "used for specifying and estimating models of linear relationships among variables" (MacCallum & Austin, 2000, p. 202). SEM can be applied both in confirmatory (testing) and exploratory (model building) modes. However, it is largely used as confirmatory technique. Here a researcher is more likely to use SEM to determine whether a certain model is valid.

A structural equation model implies a structure of the covariance matrix of the measures. Once the model's parameters have been estimated, the resulting model-implied covariance matrix can then be compared to an empirical or data-based covariance matrix. If the two matrices are consistent with one another, then the structural equation model can be considered a plausible explanation for relations between the measures.

The popularity of SEM is due to its explanatory ability and statistical efficiency for model testing with a single comprehensive method (Cheng, 2001; Hair, 2006). Many social concepts are inherently latent and cannot be observed directly and SEM has been termed particularly useful in measuring these key concepts (Westland, 2010).

SEM has the potential to test substantive theories (Kaplan, 2000). This means that SEM is more suitable to be applied to a sufficiently developed theory. And a developed theory suggests that some constructs in a model are not affecting some other constructs, which means that all variables will not load on all factors and, furthermore, certain disturbances and measurement errors will not co-vary. With the help of SEM a researcher specifies relationships between a set of variables in the form of a path diagram. Such a path diagram in SEM happens to be a simple graphical exhibit wherein uni-directional arrow indicates dependence relationship while bi-directional arrow indicates covariances, variances and correlations among the variables in the model (Hair, 2006). Apart from the arrows, there are rectangles or squares and circles or ellipses that represent the observed or *manifest variables* and unobserved or *latent variables* respectively.

To specify statistical relationships among the variables in SEM, there is always a mathematical model. To be more relevant, consider X as a manifest variable modeled as a measurement of the variable Y (latent variable), with e1 as error term. The relationship is diagramed in *Figure 1* below:

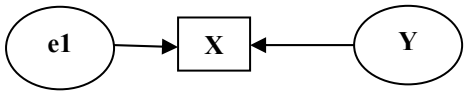

Figure 1. The measurement path between X and Y

The measurement equation for the model depicted in *Figure 1* is:

$\mathbf{X} = \lambda\mathbf{Y} + e1$

where λis the measurement strength of **X** as a measurement of **Y**.

A SEM model has two parts: the measurement part and the structural part. In the former latent variables are linked with observed variables through a restricted model that carries measurement equations for the endogenous and exogenous variables while the structural part links latent variables with each other defined by simultaneous equations.

2. Research Problem

The literature on this technique reveals that this statistical tool has widely attracted the attention of the researchers on the issue of employing various statistical techniques, on the one hand, and its usage as a practical research interpretation tool by researchers as in interdependence technique, on the other hand. However, the surprising aspect is its rare use in social sciences research. Therefore, the current research is supposed to be an introductory endeavor to look into the basic concepts and the reader be apprised of those concepts attached with this statistical technique. It is believed that this will help new researchers as well as those to have used the tool just for a single paper and want to learn those aspects as well whose need has not been felt then.

2.1 Basic Concepts in SEM

Manifest Variables: Manifest or observed variables are those variables that have numeric responses, e.g., gender or height and can be observed directly. These categories of variables in SEMs are also usually continuous and represented in a path diagram through rectangles.

2.1.1 Latent Variables

Latent or unobservable variables are those variables which have the interest of a research to measure but which cannot be observed directly. They are continuous and have infinite numbers of values and represented by circles (ellipse) on a path diagram.

2.1.2 Independent Variables

In SEM independent variables are called exogenous variables. These variables do not depend on other variables in a model and do not receive arrows in a path model.

Dependent Variables: In SEM dependent variables are called endogenous variables. They depend on other variables in a model and receive arrows in a path model.

2.1.3 Path Diagram

In SEM relationships among exogenous and endogenous variables are presented through graphic paths. This graphic presentation is called a path diagram. Herein observed variables are represented in boxes while ovals represent latent variables and disturbances-because disturbances are only estimated and cannot be measured directly-as well. For understanding of the reader a path diagram is shown in the figure 2 below.

2.1.4 Indicator Variables

In SEM latent variables are measured with the help of some observed variables. They are called indicators and are usually qualitative judgments/responses in a survey/questionnaire. The strength of each indicator with its respective latent construct is expressed as a factor loading.

2.1.5 Arrows

In a path diagram (figure 2) the links between the variables are shown with the help of arrows. A one headed arrow represents cause-and-effect or dependence relationship. Covariances or correlations between exogenous variables and between disturbances are represented by curved lines with arrowheads at both ends.

2.1.6 Measurement Model

A measurement model in SEM defines the association between the variables of interest. It provides the link between scores on a measuring instrument and the underlying constructs they are designed to measure. A measurement model is tested to validate the measurement instruments. Before testing the hypothesized relationships among the constructs of the model, the measurement must hold (Cheng, 2001). Therefore, it specifies the pattern by which each measure loads on a particular factor (Byrne, 1998).

2.1.7 Structural Model

A structural model is an explanation of the dependence among the latent variables. This explanation is based upon “a simultaneous regression of the endogenous variables in the hypothesized structural model on the predicted antecedents” (Cheng, 2001, p. 654).

2.1.8 Constrains in SEM

The main difference between simple factor analysis and SEM is that, in the former any observed variable can load on any or all factors, while, in the latter a researcher specifies which loadings and path coefficients are free to vary, and which are to be fixed (constrained). Some of the factor loadings are "constrained' or fixed to be zero. This can be seen by the absence of an arrow from one variable to another which means that the corresponding loadings in the factor matrix are fixed to zero. Similarly, a researcher is required to fix one loading to one for each factor. It is because it gives the latent factor an interpretable scale.

2.1.9 Standardized Variables

Those variables that have zero and one mean and variance respectively.

2.1.10 Disturbance

Disturbance is a corresponding term for error residual in a prediction equation. It is the set of unspecified causes of the effect variable. In SEM, usually, independent variable has a disturbance and is treated as a latent variable.

2.1.11 Specification Error

It is not possible for a researcher to include all the relevant variables in a research. This means there may be some omission and consequently a false assumption is made. E.g., when a path in a model is set to zero while that should have some value, there would be a specification error. It is not possible to specify a model without an error. However, it is a researcher's responsibility to create a model with the least amount of specification error.

2.1.12 Model Identification

It is presumed that there exists a unique solution for all of the model's parameters and such a model is termed an identified model. Here main concern is whether a unique value for each and every free parameter can be obtained from the observed data. This depends on the model choice and the specification of fixed, constrained and free parameters. However, such an identified model may not be possible. Resultantly, some of the model’s parameters may be identified. Here a researcher is required to fix the loadings of the disturbance factor to one to achieve identification (Hox & Bechger, 1998).

2.1.13 Known or Fixed and Free or Unknown Parameters

Fixed or known parameters are those whose values are shown on a path diagram along the corresponding paths

while parameters with no values on paths are free or unknown parameters. A researcher is required to estimate these unknown parameters.

2.1.14 Over-identified Model

An SEM model is a combination of known and unknown parameters. A model is termed an over-identified model if these parameters are identified and for which there are more known than free or unknown parameters.

2.1.15 Model Estimation

A procedure wherein a researcher compares the covariance matrices of the relationships between variables against the estimated covariance matrices of the best fitting model. For this purpose a number of fit statistics are employed. Each statistic has its own cut off values and a researcher interprets the data in the light of these statistics

2.1.16 Modification Indices

If there happens that a model does not fit to the data, a researcher is required to modify the model by either deleting a parameter or adding a new parameter(s) to improve the model. Software packages for SEM provide them. But a proposed index is required to be accepted only and if there is a theoretical justification for this change.

2.1.17 Goodness-of-fit Indices

Fitting a model to empirical data means solving a set of equations. By definition it is the ability of an over-identified model to reproduce the correlation or covariance matrix of the variables. There are a number of fit indices available to assess the validity of a research model. In sum, these indices are some function of the chi-square and the degrees of freedom. The purpose of these indices is to produce a goodness-of-fit that does not depend on the sample size or the distribution of the data (Hox & Bechger, 1998).

1.2 Steps in SEM

There is a four-step process in SEM. These steps are:

1.2.1Model Specification

A researcher must specify a model before he/she starts the analysis. In this specification a researcher is usually guided by a combination of theory and empirical results from the previous research (Hox & Bechger, 1998).

1.2.2 Model Identification

A specified model is then estimated with the observed data. A researcher is required to make it sure that parameters that have to be estimated have been identified. And when all parameters of a model are identified, the model is said to be identified (Bacon, 1997).

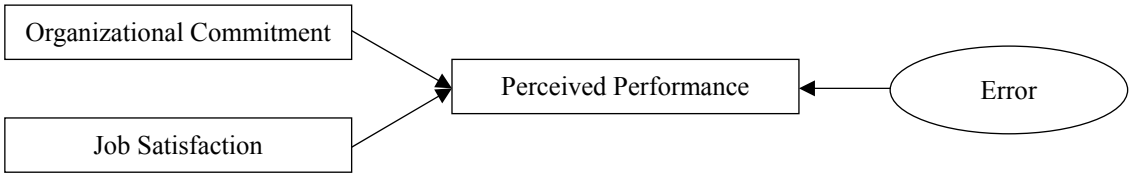

Figure 2. A modeled path diagram

1.2.3 Model Estimation

The specified model has its parameters whose values the researcher is required to estimate with the sample statistics. The most widely used method in this regard is Maximum Likelihood (ML) estimation (Hox & Bechger, 1998). It has to be kept in mind that the data is presumed to be normally distributed.

1.2.4 Model Fit

The estimated model parameters are used to predict the correlations or between measured variables and the predicted correlations or covariances are compared to the observed correlations or covariances. If the fit of the model is poor, then the model needs to be re-specified and the researcher returns to Step 1.

1.3 Statistical Assumptions in SEM

While applying any statistical technique, a researcher is required to satisfy certain assumptions which are considered something like pre-requisite. In SEM these assumptions are: multivariate normality in data, linearity in the data, large sample size, no systematic missing data, and proper model specification.

1.4 SEM-specific Software

Common software for SEM are:

✓ AMOS in SPSS

✓ Statasem

✓ SAS (software) procedures

✓ MPlus

✓ LISREL

✓ EQS

2. Historical Background

Structural Equation Modeling has a evolutionary history. It has its roots in factor analysis and *path analysis*. According to Matsueda and Press (2011) and Hoyle (2012) it was Sewall Wright, a young geneticist, who applied path analysis in the field of Genetics for causal explanation, while in the field of sociology reference to path analysis appeared in the works of Blalock (1961). It was generally believed that path analysis should exclusively be used in observational data and causality was linked with experimental data only. Side by side with causality in path analysis, interest in factor analysis was shown with the aim to have economy of description (Harman, 1960). Matsueda and Press (2011) contended that the year 1970 was a "watershed" for SEM. This year is marked by the Conference on Structural Equation Models and the publication of *Structural Equation Models in the Social Sciences*. Goldberger and Hauser (1971) and Jöreskog (1970, 1973, 1978) are great main contributors. The formers discussed many issues including identification and estimation in SEM while the latter addressed covariance analysis and introduced a computer program LISREL for empirical applications. This program dominated the field of SEM (Matsueda & Press, 2011).

The decade of 80s witnessed a major change in the form of developing of alternative indices for model estimation. The need was felt due to the sensitivity of chi-square statistics to large sample sizes. These alternatives have been divided into three categories. They are: 1) measures based on comparative fit to a baseline model, also called parsimony-based fit indices, (2) measures based on population errors of approximation, and (3) cross-validation measures. This decade also witnessed a proliferation of studies on errors in model specification. It was assumed that errors in specification can cause certain parameter problems (Kaplan, 2000).

To sum up, SEM is relatively young in the field as compared to factor and regression analysis. It has its roots in as a separate statistical tool in the papers that appeared in the late 1960s. From this one can conclude that the methodology is still in the developing phase, and even fundamental concepts are subject to challenge and revision. This rapid change is a source of excitement for some researchers and a source of frustration for others.

2.1 Criticism on SEM

With so many advantages, SEM could not be termed an absolute technique. Though it has the capability to address important substantive questions by employing simple software, the issue of fit indices is subject to severe criticism. Second, some of the assumptions like normality in data, and large sample sizes are rarely met in practice (Matsueda & Press, 2011). Third, it is more suitable in testing parsimonious models wherein restricted structural relations are taken into account. Kaplan (2000) expressed his reservations on this restricted and traditional use of SEM and argued that new methods be developed for engaging in prediction studies and evaluating predictive performance.

SEM as a statistical tool that has some short comings. MacCallum and Austin (2000) enumerate the most common of them. The first issue is that of generalizability of its findings. Like any other statistical tool its conclusions has to be kept limited to the particular sample, variables, and time frame represented in that particular study. Second, findings of SEM are subject to effects of the three important aspects (individual, measures, and occasions) of the study. A third issue is that of confirmation bias. Greenwald, Pratkanis, Leippe, and Baumgardner (1986) talk about this issue. According to them researchers using this tool are susceptible to a confirmation bias-a tilt in favor of the model under evaluation. A fourth issue is that of sample size. Though there hardly exists any rule of thumb in this regard (MacCallum, Widaman, Zhang, & Hong, 1999), findings of SEM with small sample sizes are problematic (MacCallum & Austin, 2000). Other problems include: strategies in model specification and evaluation; correlation or covariance matrices; a variety of concerns about assessment of model fit and interpretation of parameter estimates; and the difficulties associated reporting of models, method, analyses and results.

3. Areas of Application

MacCallum and Austin (2000) provides details about the usage of SEM in psychological research, both in observational and experimental studies. Though they recognized its usage more common in the observational (correlational) studies, its use in experimental studies is also of great advantage as good statistical tool. In the former, this tool can be used for both cross-sectional and longitudinal designs. A special usage of SEM is for measurement studies. This can be applied in the form of Confirmatory factor analysis (CFA) models. There is a number purposes in applying this tool. They are: for construct validation and scale refinement, multi-trait, multi-method validation, and measurement invariance (MacCallum & Austin, 2000). SEM can also be used as a tool in meta-analysis (Rounds & Tracey, 1993) as well as test-retest designs.

4. Conclusion

In research, generalizability of findings has been always a critical issue among the research community. And the choice of a relatively appropriate statistical technique has thereupon got sensitivity. Anyway, data analysis is a mean to an end. This mean is always subject to the nature of the relationship among the variables and depends on the choice of the researcher if there exist many means. However, it is very critical for a researcher that this mean be highly proper, adequate and highly helpful to achieve the end with more reliable results and convincing to the reader. There are a number of multivariate techniques, like multiple regression, factor analysis, and path analysis that help researchers drawing research conclusions. These multivariate techniques are being employed for testing single relationships between dependent and independent variables, while human and behavioral issues are much more complex wherein one dependent variable happens to be an independent variable in other dependence relationships (Cheng, 2001). It is SEM that has the potential to address these complicated issues more effectively. According to (Hair, 2006), it is SEM that has the explanatory ability and statistical efficiency for model testing with a single comprehensive method.

There are three reasons to prefer SEM over other techniques. First, SEMs utilize multiple indicator variables to measure the latent variables. Second, if one or more predictor variables are unreliable this may cause wrong coefficients or wrong signs. It may also cause model misspecification. Though separate bivariate regression may address the problem to some extent, they will not get rid of the difficulties caused by unreliable measures (Bacon, 1997). It is through SEMs that unreliability could be taken into account by trying modeling in this technique. Third, SEM has been found a powerful method for dealing with multicollinearity more effectively compared to other multivariate techniques.

References

Bacon, L. D. (1997). *Using Amos for Structural Equation Modeling in Market Research*. Lynd Bacon & Associates Limited and SPSS Incorporated.

Blalock, H. M. (1961). *Causal Inferences in Non-experimental Research*. New York: Norton.

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming.* Lawrence Erlbaum.

Cheng, E. W. L. (2001). SEM being more effective than multiple regression in parsimonious model testing for management development research. *Journal of Management Development, 20*(7), 650-667. http://dx.doi.org/10.1108/02621710110400564

Goldberger, A. S., & Hauser, R. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology, 3*(8), 1-8.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*(2), 216-229. http://dx.doi.org/10.1037/0033-295X.93.2.216

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Pearson Education. Inc.

Harman, H. H. (1960). *Modern Factor Analysis*. Chicago: University of Chicago Press.

Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review, 11*, 354-373.

Hoyle, R. H. (2012). *Handbook of Structural Equation Modeling*. Guilford Publication.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*(2), 239-251. http://dx.doi.org/10.1093/biomet/57.2.239

Jöreskog, K. G. (1973). *A general method for estimating a linear structural equation system*.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43*(4), 443-477. http://dx.doi.org/10.1007/BF02293808

Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions* (Vol. 10). Sage Newburry Park, CA.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*(1), 201-226. http://dx.doi.org/10.1146/annurev.psych.51.1.201

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99. http://dx.doi.org/10.1037/1082-989X.4.1.84

Matsueda, R. L., & Press, G. (Eds.). (2011). *Key Advances in the History of Structural Equation Modeling.* New York, NY: Guilford Press.

Rounds, J., & Tracey, T. J. (1993). Prediger's dimensional representation of Holland's RIASEC circumplex. *Journal of Applied Psychology, 78*(6), 875-890. http://dx.doi.org/10.1037/0021-9010.78.6.875

Westland, C. J. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications, 9*(6), 476-487. http://dx.doi.org/10.1016/j.elerap.2010.07.003